Corpus Linguistics for ELT

Corpus Linguistics for ELT provides a practical guide to undertaking ELT-related corpus research. Aimed at researchers, advanced undergraduate and postgraduate students of ELT and TESOL, and English language teachers, this volume:

- covers corpus research in the main areas of language study relevant to ELT: grammar, lexis, ESP, spoken grammar and discourse;
- presents a review of relevant corpus research in these areas, and discusses the implications of this research for ELT;
- suggests potential ELT-focused corpus research projects, and equips the reader with all the required tools and techniques to carry them out;
- deals with the growing area of learner corpora and direct classroom application of corpus material.

Corpus Linguistics for ELT empowers and inspires readers to carry out their own ELT corpus research, and will allow them in turn to make a significant contribution to corpus-informed ELT pedagogy.

Ivor Timmis is Reader in English Language Teaching at Leeds Beckett University. His research interests lie in the relationship between corpus linguistics and ELT, the focus of his three articles in ELT Journal. He is also working on a historical spoken corpus with data from the late 1930s in his home town, Bolton.
Routledge Corpus Linguistics Guides provide accessible and practical introductions to using corpus linguistic methods in key sub-fields within linguistics. Corpus linguistics is one of the most dynamic and rapidly developing areas in the field of language studies, and use of corpora is an important part of modern linguistic research. Books in this series provide the ideal guide for students and researchers using corpus data for research and study in a variety of subject areas.

Ronald Carter is Research Professor of Modern English Language in the School of English at the University of Nottingham, UK. He is the co-series editor of the Routledge Applied Linguistics, Routledge Introductions to Applied Linguistics and Routledge English Language Introductions series.

Michael McCarthy is Emeritus Professor of Applied Linguistics at the University of Nottingham, UK, Adjunct Professor of Applied Linguistics at the University of Limerick, Ireland and Visiting Professor in Applied Linguistics at Newcastle University, UK. He is co-editor of the Routledge Handbook of Corpus Linguistics and editor of the Routledge Domains of Discourse series.

OTHER TITLES IN THIS SERIES

**Corpus Linguistics for Grammar**
*Christian Jones and Daniel Waller*

**Corpus Linguistics for ELT**
*Ivor Timmis*

**Corpus Linguistics for Discourse Analysis**
*Michael Handford*

**Corpus Linguistics for Sociolinguistics**
*Bróna Murphy*

**Corpus Linguistics for the Social Sciences**
*Tony McEnery, Amanda Potts, Vaclav Brezina and Andrew Hardie*
For Mum, Dad, Harriet and Robbie
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of figures</td>
<td>xi</td>
</tr>
<tr>
<td>List of tables</td>
<td>xiii</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>xv</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Aims</td>
<td>1</td>
</tr>
<tr>
<td>What is a corpus?</td>
<td>2</td>
</tr>
<tr>
<td>Defining a corpus</td>
<td>2</td>
</tr>
<tr>
<td>Types of corpus</td>
<td>3</td>
</tr>
<tr>
<td>What can we do with a corpus?</td>
<td>4</td>
</tr>
<tr>
<td>Questions corpora can answer – quantitative analysis</td>
<td>4</td>
</tr>
<tr>
<td>Qualitative corpus analysis</td>
<td>6</td>
</tr>
<tr>
<td>Corpora and ELT</td>
<td>7</td>
</tr>
<tr>
<td>Corpora and ELT: In need of relationship counselling?</td>
<td>7</td>
</tr>
<tr>
<td>What do corpora have to offer ELT?</td>
<td>8</td>
</tr>
<tr>
<td>2 Building a corpus</td>
<td>14</td>
</tr>
<tr>
<td>Why build your own corpus?</td>
<td>14</td>
</tr>
<tr>
<td>Building a corpus: A basic guide</td>
<td>14</td>
</tr>
<tr>
<td>Analysing corpus data</td>
<td>17</td>
</tr>
<tr>
<td>Word frequency counts</td>
<td>17</td>
</tr>
<tr>
<td>Concordancing</td>
<td>18</td>
</tr>
<tr>
<td>Collocations</td>
<td>19</td>
</tr>
<tr>
<td>3 Corpora and lexis</td>
<td>22</td>
</tr>
<tr>
<td>Introduction</td>
<td>22</td>
</tr>
<tr>
<td>Describing the lexicon</td>
<td>24</td>
</tr>
<tr>
<td>Terminology</td>
<td>24</td>
</tr>
<tr>
<td>Applying our terminology</td>
<td>29</td>
</tr>
</tbody>
</table>
4 Corpus research and grammar
Introduction 57
Grammar and frequency 58
Frequency findings from corpus research 58
Corpus research into frequency: Pedagogic implications 60
Corpus perspectives on grammatical description 62
Corpus perspectives on reported speech 62
Conditionals, corpora and pedagogy 64
Not such a simple word though 68
Grammar and lexis: Connections 71
The lexis and grammar connection 71
The grammar and lexis connection 74
The grammar/lexis and lexis/grammar connection:
Implications for pedagogy 75
Conclusion 77

5 Spoken corpus research
Introduction 81
Examples of spoken corpora 82
Spoken corpus research and lexis 84
Spoken word frequency 84
Frequency of collocations and lexical chunks 88
Spoken corpus research and grammar 91
Spoken discourse 96
Discourse sequences 99
Spoken corpus research and pragmatics 101
Syllabus, materials and methodology 103
Syllabus selection criteria 103
Syllabus and sequence 108
Spoken language and methodology 110
Spoken language and materials 113
Conclusion 115
6  Corpora and the classroom 119

Introduction 119
Learner corpora 119
  What is a learner corpus? 119
  Design criteria for a learner corpus 120
  Learner corpus research: Theory and practice 125
  Applications of learner corpus research 127
Teaching-oriented corpora 128
  The need for teaching-oriented corpora 128
  Examples of teaching-oriented corpora 129
  Improvising corpora for the classroom 131
Data-Driven Learning (DDL) 133
  What is data-driven learning? 133
  What is the rationale for DDL? 135
  Reservations about DDL 138
  DDL in perspective 139
  Final word on DDL 141

7  Corpora and ESP 146

Introduction 146
Types of EAP corpus 147
Examples of EAP corpora 148
English for Academic Purposes corpus research 151
  English for Academic Purposes corpora and lexis 151
  English for Academic Purposes corpora and genre analysis 159
  Engineering corpora 161
  Corpora of Business English 166
  Learner corpora and ESP 171
  Corpus findings and the ESP syllabus 172
  English for Specific Purposes corpora and DDL 174
  From corpus to materials 178
Conclusion 180

8  Corpora in perspective 183

Introduction 183
Limitations of corpora 183
Corpora, the native speaker and English as a Lingua Franca 184
  Native speakers in perspective 184
  English as a Lingua Franca research 185
  Classroom options for models of English 189
## List of figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>BNC KWIC concordance: <em>responsible</em></td>
<td>11</td>
</tr>
<tr>
<td>2.1</td>
<td>COCA word frequency list</td>
<td>18</td>
</tr>
<tr>
<td>2.2</td>
<td>COCA KWIC search interface: <em>dulcet</em></td>
<td>19</td>
</tr>
<tr>
<td>2.3</td>
<td>BNC KWIC concordance: <em>dulcet</em></td>
<td>19</td>
</tr>
<tr>
<td>2.4</td>
<td>BNC collocation search: <em>dulcet</em></td>
<td>20</td>
</tr>
<tr>
<td>3.1</td>
<td>BNC search interface</td>
<td>25</td>
</tr>
<tr>
<td>3.2</td>
<td>BNC search: <em>carry out</em></td>
<td>25</td>
</tr>
<tr>
<td>3.3</td>
<td>N-gram search 1 using antconc software</td>
<td>28</td>
</tr>
<tr>
<td>3.4</td>
<td>BNC search: <em>carry out</em> + noun</td>
<td>30</td>
</tr>
<tr>
<td>3.5</td>
<td>N-gram search 2 using antconc software</td>
<td>33</td>
</tr>
<tr>
<td>3.6</td>
<td>BNC KWIC concordance: <em>sheer</em></td>
<td>37</td>
</tr>
<tr>
<td>3.7</td>
<td>BNC collocation search: <em>come up with</em></td>
<td>38</td>
</tr>
<tr>
<td>3.8</td>
<td>BNC collocation search: <em>happen</em></td>
<td>38</td>
</tr>
<tr>
<td>3.9</td>
<td>BNC collocation search: <em>broke out</em></td>
<td>39</td>
</tr>
<tr>
<td>3.10</td>
<td>BNC collocation search: <em>mystery word</em></td>
<td>40</td>
</tr>
<tr>
<td>3.11</td>
<td>Academic word frequency highlighter</td>
<td>44</td>
</tr>
<tr>
<td>3.12</td>
<td>BNC concordance lines: <em>thing</em></td>
<td>53</td>
</tr>
<tr>
<td>4.1</td>
<td>BNC (spoken) concordance lines: <em>though</em></td>
<td>69</td>
</tr>
<tr>
<td>4.2</td>
<td>BNC search: <em>What I [verb] about</em></td>
<td>73</td>
</tr>
<tr>
<td>4.3</td>
<td>BNC collocation search: <em>bordered on</em></td>
<td>74</td>
</tr>
<tr>
<td>4.4</td>
<td>BNC search: <em>has been</em></td>
<td>78</td>
</tr>
<tr>
<td>5.1</td>
<td>BNC concordance lines: <em>nightmare</em></td>
<td>105</td>
</tr>
<tr>
<td>6.1</td>
<td>BNC concordance: <em>persuade</em></td>
<td>133</td>
</tr>
<tr>
<td>6.2</td>
<td>BNC concordance: <em>convince</em></td>
<td>134</td>
</tr>
<tr>
<td>6.3</td>
<td>Academic word highlighter gap-fill</td>
<td>137</td>
</tr>
<tr>
<td>7.1</td>
<td>Hong Kong Engineering Corpus search interface</td>
<td>164</td>
</tr>
<tr>
<td>7.2</td>
<td>Hong Kong Engineering Corpus ConCgram results</td>
<td>165</td>
</tr>
<tr>
<td>7.3</td>
<td>ConCgram: <em>expenditure</em></td>
<td>166</td>
</tr>
<tr>
<td>7.4</td>
<td>Concordance: <em>solution</em></td>
<td>179</td>
</tr>
</tbody>
</table>
This page intentionally left blank
## List of tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Node word and collocates</td>
<td>48</td>
</tr>
<tr>
<td>3.2</td>
<td>Lexical phrase frequency list (adapted from Martinez and Schmitt 2012)</td>
<td>50</td>
</tr>
<tr>
<td>4.1</td>
<td>Conditionals typology (adapted from Maule 1988)</td>
<td>66</td>
</tr>
<tr>
<td>5.1</td>
<td>Spoken language ‘syllabus’</td>
<td>109</td>
</tr>
<tr>
<td>7.1</td>
<td>Academic word list (sublist 1) (Coxhead 2000)</td>
<td>153</td>
</tr>
<tr>
<td>7.2</td>
<td>Academic bundles frequency list (adapted from Hyland 2008)</td>
<td>158</td>
</tr>
<tr>
<td>7.3</td>
<td>Common phraseology used in the moves in the JPA Corpus</td>
<td>160</td>
</tr>
<tr>
<td>(adapted from Hsieh and Liou 2008)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.4</td>
<td>Nelson’s key words in Business English</td>
<td>167</td>
</tr>
<tr>
<td>7.5</td>
<td>Nelson’s most common 2- and 3-word phrases in Business English</td>
<td>170</td>
</tr>
</tbody>
</table>
This page intentionally left blank
I would like to thank Ronald Carter and Michael McCarthy for inspiring my interest in corpus linguistics and for supporting my efforts in the field over many years. The IVACS group of researchers has been a much appreciated source of ideas, informed feedback and convivial company. Thanks also to Jane Templeton at Leeds University for a constructive and engaging reader’s eye view on some of the chapters.
Chapter 1

Introduction

Aims

The challenge of fostering a fruitful relationship between corpus linguistics and ELT was clearly set out by Conrad (2000: 556):

Corpus grammarians must strive to reach more audiences that include teachers and must emphasize concrete pedagogical applications … In fact, the strongest force for change could be a new generation of ESL teachers who were introduced to corpus-based research in their training programs [and] have practiced conducting their own corpus investigations and designing materials based on corpus research.

Indeed, this comment by Conrad encapsulates the main aim of this book: to help move corpus linguistics from what Römer (2012) terms its ‘minority sport’ status in language teaching to a point where the ability to carry out and interpret corpus research is seen as a normal part of an English language teacher’s repertoire. Familiarity with corpus research and practice should be a standard part of an English language teacher’s toolkit, I would argue, because most people in ELT will at some time have had thoughts like these:

- How many words do my learners need to learn?
- Why is everyone talking about lexical chunks and collocations?
- Do my students really need this grammar point?
- Which words should I use to exemplify this structure?
- Am I teaching my learners language they will need to use when they speak the language?
- Does the grammar explanation in the coursebook really reflect how we use this structure?
- What vocabulary do my English for dentistry students need to get their teeth into?
If you have had questions like these, this book is designed to help you to answer them by consulting corpora and corpus-informed literature. It is also designed to help you to generate and investigate similar questions. It is, however, important to keep corpora in perspective throughout this book. The argument presented here is that corpora are a resource and a reference source and, as is the case with all resources, pedagogic judgement is vitally important in determining how and when they are deployed to best effect.

The book does not assume prior knowledge or experience of corpus research; nor does it assume any technical expertise. Technophobes can relax: contemporary corpus interfaces and corpus software are user-friendly and often include tutorial packages. The tasks in this book will help to familiarise readers with publicly available user-friendly corpora such as the British National Corpus hosted at http://corpus.byu.edu/bnc/

And if you know how to save a document, you are, as we shall see in the next chapter, well on the way to being able to compile your own corpus for teaching purposes; and then things get really interesting.

What is a corpus?

Defining a corpus

If you are reading this book, you probably know what a corpus is, but it is useful to draw out some key points from definitions in the literature to be sure that we have a shared understanding. Brazil (1995: 24) defines a corpus as ‘a collection of used language’, explaining that ‘used language’ is ‘language which has occurred under circumstances in which the speaker was known to be doing something more than demonstrate the way the system works’. This definition is useful in that it focuses on the fact that language in a corpus is naturally occurring. We need to note, however, that a corpus is not just a collection of naturally occurring language in the form of isolated words or sentences randomly collected; it consists of spoken and/or written texts (the word ‘text’ in corpus linguistics is used to refer to both spoken and written language). And the collection of texts also has to be purposeful: ‘A corpus is not simply a collection of texts. Rather a corpus seeks to represent a language or some part of a language’ (Biber, Conrad and Reppen 1998: 246). In practice, as McEnery and Wilson (1996) note, in contemporary usage a corpus almost always refers to texts collected in machine-readable form, i.e. electronic texts which can be automatically analysed with software packages. For our purposes, it is important to note that while ‘big-name’ corpora such as the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA) consist of hundreds of millions of words, size is not an absolute criterion for corpus design: size is a question of fitness for purpose. O’Keeffe, McCarthy and Carter (2007: 4) stress that the design of the corpus is more important than the size:
For corpora of spoken language, anything over a million words is considered to be large; for written corpora, anything below five million is considered quite small. In terms of suitability, however, it is often the design of a corpus as opposed to its size which is the determining factor.

It is the design of a corpus which will ensure that it represents what it seeks to represent. Design issues include demographic factors such as gender, age and social class, as well as questions of the genres and contexts of the language included in the corpus. Even a very large corpus such as the BNC self-evidently does not tell us how English is used in the USA, in India, or as a lingua franca between non-native speakers.

**Types of corpus**

It is important to be aware of the range of corpora available (see Appendix 2 for a fuller list). While large general corpora such as BNC and COCA have both written and spoken components, many corpora are either written or spoken. The five million word CANCODE (Cambridge and Nottingham Corpus of Discourse English) is a well-known spoken corpus often cited in ELT studies. There are also English for Specific Purposes corpora, e.g. MICASE (Michigan Corpus of Academic Spoken English); CANBEC (Cambridge and Nottingham Business English Corpus), and the Hong Kong Engineering corpus. For ELT purposes, corpora of non-native English are important, e.g. VOICE\(^1\) (Vienna–Oxford International Corpus of English), a spoken corpus of English used as a Lingua Franca (ELF). Learner corpora are a specific type of non-native corpus, self-evidently containing data produced by learners of English, e.g. ICLE (International Corpus of Learner English) which ‘contains argumentative essays written by higher intermediate to advanced learners of English from several mother tongue backgrounds’ (http://www.uc louvain.be/en-cecl-icle.html). We need to consider one further type of corpus: a pedagogic corpus or, to use Leech’s (1997) term, a teaching-oriented corpus. A pedagogic corpus is one that has been compiled specifically for language teaching purposes. An interesting suggestion for ‘pedagogic corpora’ has been made by Willis (2003), who proposes a pedagogic corpus is made up of the texts already used by the learners in class, which is then exploited for the study of particular language features. The advantage of such corpora, Willis (2003) argues, is that learners will already be familiar with the co-text, i.e. the text immediately surrounding the target feature, as they will previously have studied the whole text in class. Similarly, Römer (2006) has suggested that coursebooks themselves can be made into corpora so that ‘coursebook English’ can be compared with ‘real English’. The SACODEYL (System Aided Compilation and Distribution of European Youth Language) corpus could also be seen as a pedagogic corpus, though it was not compiled from learning materials; it was deliberately constructed for language learning purposes, as described below on the SACODEYL website: ‘The [SACODEYL] corpora are based on structured
video interviews with pupils between 13 and 18 years of age. The interviews have been annotated and enriched for language learning purposes. ’ http://sacodeyl.inf.um.es/sacodeyl-search2/

While SACODEYL might not be the most transparent project title, it has the significant benefit of being free to access and providing online guidance on how to use it.

**Corpus Search**

Visit the four websites below and consider which you might find most useful for your teaching, research or studies:

- http://corpus.byu.edu/bnc/
- http://sacodeyl.inf.um.es/sacodeyl-search2/
- http://www.univie.ac.at/voice/

**What can we do with a corpus?**

**Questions corpora can answer – quantitative analysis**

Though corpus linguistics has come to be seen as a domain of applied linguistics in its own right, it will be useful for our purposes to view it also as a *methodology* through which various domains of applied linguistics can be investigated, e.g. grammar, lexis, discourse, pragmatics, SLA (second language acquisition). Corpora are most often associated with quantitative research as frequency information can be generated with striking ease. The most basic kinds of frequency question we can ask are:

1. What are the most frequent words in our corpus, i.e. rank order?
2. How many instances of a given word are there in the corpus, i.e. raw frequency?
3. What percentage of the total number of tokens in the corpus does the raw frequency represent, i.e. relative frequency?
4. What are the most frequent collocations of a given word in our corpus?
5. What are the most frequent phrases of a given length (e.g. 2-word phrases, 3-word phrases, 4-word phrases and so on)?
6. What are the most frequent grammatical structures in our corpus?

Each of these questions may be applied with a more specific focus, but we will take word frequency as an example:
1. What are the most frequent words used in a given component of the corpus, e.g. academic or business or technical English?

2. What are the most frequent words used by a particular demographic group of people, e.g. women, people under 30, people of a given social class or from a given region?

3. What are the most frequent words used in a particular kind of text, e.g. scientific articles?

4. What are the most frequent words in a given genre, e.g. self-descriptions on internet dating sites?

These questions do not exhaust the possibilities, but give some idea of the range of questions which can be asked of a corpus. It is crucial to note, however, that the kind of question which can be investigated depends on the composition of the corpus and the information which has been encoded about the speakers/writers, the setting, the context and so on. This encoded information is known as metadata, i.e. data about the linguistic data.

While frequency information is straightforward to obtain, it can yield interesting insights about the nature of language, as demonstrated in the task below.

**Corpus Question**  
(see Appendix 1, Commentary 1 for answers)

a. Word frequency in spoken language

*Please put the following words in order of their frequency in spoken British English.*

Think; House; Well; Money; Good; Just; Right; Weather

b. Collocation frequency in spoken language (Shin and Nation 2008)

*Please put the following collocations in order of their frequency in spoken British English.*

thank you; I think; you know; very nice; a bit; in fact; I suppose; this morning

Even from this short activity, we can see how frequency alone can provide more general insights into language. The frequency of *well*, *just* and *right*, for example, shows the importance in spoken language of discourse markers (a point to which we return in Chapter 5), especially discourse markers with an interactive role (O’Keeffe, McCarthy and Carter 2007). The presence of the discourse marker *you know* at the top of the collocation frequency list underlines the importance of interactive concerns, while the collocations *I think* and *a bit* both illustrate the centrality of ‘hedging’ in spoken language, i.e. down-toning an utterance to make it sound less definite.
We need also to consider comparative frequencies between corpora. For this purpose, it is useful if frequency counts are normalised, i.e. expressed in terms of frequency per 1,000 words or per 10,000 words, for example. If you are working with a (small) specific corpus, for example, you may wish to ascertain whether a given finding is unique to that corpus or simply reflects general usage. In a small corpus of English spoken in pubs (the Bolton Corpus, Timmis 2010) that I was working with, for example, the word bloody was highly frequent, but that turned out to be the effect of the context in which the conversations were recorded: it was far less frequent in the BNC. For this kind of comparison, you can use the keyword method to determine if there is a statistically significant difference in frequency between a given word in your corpus and the same word in a general corpus, known as a reference corpus for this kind of operation.

**Qualitative corpus analysis**

Many interesting research questions, however, are not amenable to purely quantitative corpus research and have to be researched qualitatively, as Timmis (2013: 463) points out:

Automatic corpus analysis will not tell us, for example, which of the various meanings of the word ‘tip’ is most frequent, which of the various uses of the present perfect is most frequent, or how often the word ‘marvellous’ is used sarcastically. We need manual, qualitative reading of corpus data to supply us with this kind of information.

We should note here that Timmis above refers to different kinds of research focus: lexical, grammatical and pragmatic corpus research.

Let us take two specific examples of qualitative corpus research of potential relevance to language teaching:

1. McCarthy (1998) reports that in the CANCODE corpus 124 out of 139 instances of the ‘Get passive’ (e.g. ‘I got caught in traffic’) are adversative, i.e. they refer to an action viewed negatively by the speaker. This kind of finding is only possible if the researcher reads each example and interprets whether the example is adversative or not.

2. Conrad (2004) argues that in conversation the discourse marker though is often used to ‘soften disagreement’ between speakers, e.g. (attested example)
   
   S1: I’m sick of all this paperwork.
   
   S2: We do need to keep track of progress though.

A corpus is not capable of automatically recognising the pragmatic behaviour illustrated in the example above, i.e. softening behaviour; or at least it is not capable of doing that yet …
One might, it is true, arrive at a hypothesis in relation to these two points – the ‘get passive’ and the interactive use of *though* – intuitively, but the corpus offers important evidence which the researcher can use to confirm, refute or refine the hypothesis.

**Corpora and ELT**

*Corpora and ELT: In need of relationship counselling?*

We have briefly outlined the potential for corpora to contribute to linguistic research in general above; we now need to consider its specific potential in relation to ELT. A good starting point is Conrad’s (2000: 548) question below:

> Will corpus linguistics revolutionize grammar teaching in the 21st century?

The simple answer to this question is, ‘on the evidence so far, “no”’. We need, however, to consider this question more carefully as it goes to the heart of the purpose of this book. Firstly, we should note that there is no reason why the question should necessarily be limited to revolutionising grammar teaching given that corpus research embraces other areas of linguistic study highly relevant to ELT such as lexis, discourse and pragmatics. Secondly, and most importantly, we need to consider why the question should be expressed in such dramatic terms, i.e. ‘revolutionise’. The answer to this question lies in the claims which have been made for the power of corpus linguistics. Sinclair (1991), for example, likened the value of corpora for linguistics to the value of the telescope for astronomy. The basic argument for the application of corpus linguistics to language teaching is, then, that the quantity and quality of evidence about language use available through corpora can lead (and have led) to sounder and more comprehensive descriptions of language than previous research methods allowed. However, there is a more radical argument (e.g. Sinclair 1991; Hunston 2002) for applying the insights from corpus linguistics to language teaching: the cumulative effect, it has been argued, of these individual descriptive insights has been to present a picture of language itself which is significantly different from long-held traditional views which have traditionally seen grammar as primary in communication and grammar and lexis as distinct domains. Cook (1998: 57), for example, summarises the more general insights gained into the nature of language from corpus evidence:

- Actual language use is not so much a matter of applying abstract grammatical rules in combination with lexical items, but more a question of collocation.
- Some utterances which are grammatically possible do not occur, while others occur with disproportionate frequency.
- The domains of grammar and lexis are less distinct and more mutually dependent than previously thought.
As O’Keeffe, McCarthy and Carter (2007: 21) note: ‘The contribution of corpus linguistics … to the description of the language we teach is difficult to dispute’. Even if we accept this claim, however, there is, as O’Keeffe, McCarthy and Carter (2007: 21) indeed acknowledge, a need to sound a cautionary note: language description is only one aspect of language teaching, which is a highly complex activity taking place in different contexts, for different purposes and involving teachers and learners with different educational and cultural expectations, teaching and learning styles. Teaching is self-evidently not simply a matter of transferring descriptive insights about language to learners, however accurate and comprehensive these descriptions may be and whatever their source. And, as we shall discuss later, there are also significant questions about whose English is being described in corpora and what authority the particular variety of English captured in a corpus should carry in the complex world of contemporary English language use.

This book does not propose, then, to ‘revolutionise’ language teaching through corpus linguistics. It does, however, seek to show how the long proclaimed potential of corpora can significantly contribute to the evolution of language teaching. Ideally, it would take us to a point where we stop talking about the potential of corpora to be applied to ELT, which has been proclaimed for many years, to a point where the ability to carry out corpus research and/or interpret the significance of corpus findings for the classroom is regarded as the norm for ELT practitioners. At the moment, this does not seem to be the case: a number of commentators (e.g. Römer 2006; Biber and Conrad 2010; Burton 2012) report that only a minority of students, teachers and materials writers directly engage with or in corpus research in practice, despite the growing number of both general and specialised corpora available, as Römer (2006:124) emphasises:

Despite the obvious and recognised strengths of corpus use in a pedagogical context, e.g. that corpora highlight what lexical items and collocations are typical in the language, and that they provide us with large amounts of natural language examples … it seems that there is still a strong resistance towards corpora from the side of students, teachers, and materials writers.

It is interesting that Römer (2006) refers above to the ‘obvious and recognised strengths of corpus use in a pedagogical context’, while at the same time speaking of ‘resistance’ towards corpora from those most closely involved in the pedagogic context: why should people resist what is potentially in their interest? Before we can answer that question, we will need to critically analyse the contribution corpora can make to language teaching.

**What do corpora have to offer ELT?**

Let us turn now in more detail to this proclaimed potential of corpora: Leech (1997) spoke optimistically about the potential of corpora to contribute to language teaching in three ways:
The indirect use of corpora in teaching, i.e. corpora can be used to inform ELT reference works such as dictionaries and grammars, to inform the content of ELT materials and syllabuses, and to inform test design.

The use of corpora to inform reference works is now a given. The CoBuild series of dictionaries and reference books based on data from the Bank of English corpus were the forerunner in this field and now, as McCarten and McCarthy (2010) point out, an ELT reference book which did not claim to be corpus-informed would be conspicuous. However, the relationship between corpus research and ELT materials and syllabus design is nowhere near as strong as the relationship between corpus research and ELT reference books. Nevertheless, we should note that McCarten and McCarthy (2010: 13) do point to the influence of corpora on a small number of coursebooks, e.g.

- **face2face** (Redston and Cunningham 2005), which is informed by Cambridge International Corpus + Cambridge Learner Corpus.
- **Objective First Certificate** (Capel and Sharp 2008), which is informed by the Cambridge Learner Corpus (50,000 written examination papers – substantially error-coded).
- **Touchstone** (McCarthy, McCarten and Sandiford 2005, 2006) where ‘extensive use was made of a corpus to inform not only the grammatical and lexical syllabuses, primarily through frequency lists and concordances, but also the topics and methodology…’ McCartney and McCarthy (2010: 13).

**Corpus Question**

If possible, access one or more of the three coursebooks mentioned above. Examine the syllabus/contents pages and then one unit in detail. Can you see evidence of the influence of corpora? If so, where and how?

While the influence of corpora may be evident in the coursebooks mentioned above, the title of Burton’s (2012) article, ‘Corpora and coursebooks: destined to be strangers forever’, suggests a rather distant and uncomfortable relationship between corpus research and materials. Indeed, the fact that McCarten and McCarthy (2010) were able to name the books influenced by corpora suggests that this is the exception rather than the rule. An important aim of this book, then, is to foster a more productive relationship between corpora and ELT by reviewing the findings of ELT-related corpus research, suggesting avenues for further research, and considering how judicious use of corpus research could contribute positively to ELT practice.
2 The direct use of corpora in teaching, i.e. using corpus data in the classroom and enabling learners to access corpora for autonomous study.

The direct use of corpora or corpus data in the classroom is usually known as data-driven learning (DDL) and most commonly associated with the pioneering work of Tim Johns (1991) in this area. In DDL, learners may, for example, be given a concordance printout containing numerous instances of a particular word or phrase presented with brief co-text; they are then asked to make observations on its meaning, use and grammatical properties based on the evidence before them. Below is a very, very simple example – we will consider more sophisticated examples in Chapter 6. If, like me, you have spent many years correcting students who say ‘responsible of’, you might ask them to carry out a corpus search, perhaps using the free-to-access BNC interface at http://corpus.byu.edu/bnc/. In this case, the sheer weight of visual evidence can make an impact (see Figure 1.1).

DDL, as we shall see in Chapter 6, places great emphasis on learners discovering language to the extent that it may challenge the beliefs of both learners and teachers. We will consider various ways in which this kind of learning might be applied in the classroom.

3 Further teaching-oriented corpus development, i.e. developing corpora of learner English, corpora of L1 language development and corpora of English for Specific Purposes.

Given that, as we shall see in later chapters, corpora of learner English, L1 language development and ESP already exist, building your own corpus may seem unnecessary and, as O’Keeffe and Farr (2003) remark, it can also seem to be a daunting task. However, as we shall see in Chapter 2, it does not have to be daunting: a small but useful written corpus can be built in a matter of hours. The question, however, is why we would want to do this? In our case, it is probably to be able to meet the precise needs of our learners by, for example, providing a corpus which addresses their specialism and/or which is suitable for their level of English. Let us take a look at an example from my experience. In our institution we ran a course for Spanish teachers of music to help them to prepare to deliver CLIL (Content and Language Integrated Learning) music lessons, i.e. they were to begin teaching music to their Spanish pupils through the medium of English. How could they best be prepared for this new challenge? One way, we decided, was to compile a small corpus of ‘music teaching English’. Ideally, this would have involved recording and transcribing music lessons in English schools, but this was too time-consuming and expensive for us at that time. We opted for what Tribble (1997) calls a ‘quick and dirty corpus’, also referred to as ad hoc corpora (Römer 2012), i.e. small, informally produced corpora designed for a specific teaching/learning purpose. Compiling the music corpus involved downloading from a website the comments of music examiners on the performance of English music students in the GCSE music exam.
Figure 1.1  BNC KWIC concordance: responsible
• It was quick: in about 45 minutes we had the raw material for a 10,000 word corpus of ‘English used to assess musical performance’.
• It was dirty in the sense that it was clearly imperfect: we had no knowledge of who had written the reports or what their brief was and it was not English used to teach music in the classroom.
• It was, we think, effective for our immediate purposes: we used it on the course and it was well received.

Finally in this chapter, a cautionary tale: it was once said of an English professional footballer that ‘he has a great future behind him’ – his long proclaimed youthful promise never came to fruition. We need to make sure that the same is not said of the relationship between corpus linguistics and ELT: there is a time to move beyond talking only about potential.

Throughout the book you will find three kinds of activity:

**Corpus Search**: Usually a visit to a corpus site and a guided search.
**Corpus Question**: This usually involves analysis and discussion of data.
**Discussion**: This usually involves discussion of classroom implications.

**Note**

1 ‘… a sizeable, computer-readable corpus of English as it is spoken by this non-native speaking majority of users in different contexts. These speakers use English successfully on a daily basis all over the world, in their personal, professional or academic lives. We therefore see them primarily not as language learners but as language users in their own right.’ http://www.univie.ac.at/voice/

**References**

Why build your own corpus?

We saw in the example of the music corpus in Chapter 1 that there is sometimes a case for building your own corpus, even in a world where ready-made corpora are increasing in size, variety and availability. There are, indeed, several possible reasons why you might want to build your own corpus:

- To inform yourself about the language typically used in a genre or field you are currently teaching, e.g. writing business reports, English for tourism, English for science.
- To inform the syllabus for a genre or field you are currently teaching.
- As a direct source of materials, i.e. DDL (data-driven learning).
- To provide a resource for learners to use autonomously which is appropriate in terms of topic, level and accessibility.
- To analyse your learners’ language.

It is important to note that these objectives are not mutually exclusive and a suitably designed corpus may encompass one or more of these aims.

Building a corpus: A basic guide

The following remark (O’Keeffe, McCarthy and Carter 2007: 5) should be re-assuring to readers with no prior experience of building corpora:

A basic language corpus can be assembled from spoken and written texts and can be used with commercially available corpus software such as Wordsmith Tools (Scott 1999) and Monoconc Pro (2000) which any average home computer user can manipulate with relative ease.

We should remind ourselves here that ‘text’ in corpus discussions is used to refer to spoken language as well as written language, noting, however, that while it is normally self-evident where a written text begins and ends, this is not the case
with spoken texts, particularly in relation to casual conversation. We have already noted that the collection of texts for a corpus needs to be purposeful. The fundamental question is:

- What kind of language use do you want your corpus to represent?

Choices depend on a range of demographic, generic and contextual factors, i.e.:

- Whose language use are you trying to represent?
- What kind of genre(s) are you trying to represent?
- What kind of context are you trying to represent?
- Who are the potential users of your corpus?

The possibilities for corpus construction are almost infinite. Here are just a few examples of corpora compiled in an educational context:

- The language used by teachers in meetings (Vaughan 2007).
- The language used by learners in Cambridge written examinations (Cambridge English Profile Corpus) http://www.englishprofile.org/index.php/corpus
- The language used in teaching practice feedback (Farr 2008).
- The language used by ‘teachers telling anecdotes’ (Timmis 2010). (We refer to this corpus a number of times later in the book as the TTT Corpus – Teachers Telling Tales.)

The crucial decision as to what it is you want to represent goes hand-in-hand with the question of whether you need spoken or written data or both. It is generally easier to compile written corpora because of the ready availability of written texts in electronic form: the internet is a rich source, though due attention has to be paid to copyright. Spoken corpora are generally more difficult to compile because there is the need both to make and to transcribe the recordings, which is time-consuming. O’Keeffe, McCarthy and Carter (2007), for example, estimate that one hour of recorded speech produces 12–15,000 words of data and this typically takes two days to transcribe. Transcription, as de Cock (2010) points out can be ‘broad’ or ‘narrow’:

A distinction is often made between two extreme types of transcriptions: ‘broad’ and ‘narrow’ or ‘fine-textured’ transcriptions … While broad transcriptions provide only little information over and above the verbatim record of what is said, narrow or fine-textured transcriptions provide considerable detail regarding aspects such as voice quality, intonation, stress and other phonetic/phonemic details of pronunciation.

How broad or narrow the transcription needs to be will be determined, de Cock (2010) observes, by the researcher’s purposes and the resources available. In our
case, as our concerns are pedagogic, we are likely to be content with quite broad
descriptions.

Once you have decided what it is you want to represent, the next stage is to
collect texts and store them electronically. A number of points need to be
considered in relation to text selection:

- How big is your corpus going to be?
- How many texts are you going to use from different categories?
- Are you going to use complete texts or samples from texts?

While we have stressed the need for a corpus to be representative, we should, given
that our purposes are ultimately related to the practical business of ELT, take note
that Sinclair (2005: 81) cautions against perfectionism in corpus development:

> It is important to avoid perfectionism in corpus building. It is an inexact
> science, and no-one knows what an ideal corpus would be like … Until then
> compilers make the best corpus they can in the circumstances, and their
> proper stance is to be detailed and honest about the contents. From their
> description of the corpus, the research community can judge how far to trust
> their results, and future users of the same corpus can estimate its reliability
> for their purposes.

The texts may also need to be ‘cleaned’, e.g. with internet texts you may need to
remove photos, advertisements, hyperlinks and so on. It is best to save texts in
plain text format to facilitate analysis by different types of software. Sinclair
(2005) advises that each text should be given an identification code linked to a
database with information about that text. Corpus compilers may also wish to
encode other information about the text (metadata), e.g. speakers, date, source of
text etc. This kind of information is encoded in mark-up format, which is defined
by McEnery, Xiao and Tono (2006: 22) as ‘a system of standard codes inserted
into a document stored in electronic form to provide information about the
<text>text</text> itself and govern formatting, printing or other processing’. Two widely used
mark-up schemes are TEI (Text Encoding Initiative – www.tei-c.org) and CES
(Corpus Encoding Standard – http://www.xces.org/). Corpus texts are coded in a
mark-up format so that the information about the text can be kept separate from
the corpus data itself when the software ‘reads’ the material. To take a specific
example, in a simple spoken corpus I prepared, I had labelled each conversation
in the corpus ‘dialogue 1’, ‘dialogue 2’ etc., but without using mark-up format for
these labels. This meant that the word ‘dialogue’ appeared high in the word
frequency list in my spoken corpus even though none of the speakers had actually
used the word. If I had used mark-up format for the word ‘dialogue’ the software
would not have read it when I carried out searches. Mark-up format also facilitates
more refined searches: you may, for example, want to focus your corpus search on
a specific kind of text or on a specific demographic group of speakers.
Large-scale corpora designed for public use are often grammatically ‘tagged’, i.e. each word in the corpus is given a grammatical label (tag) to facilitate grammatical searches. This can be done with online tools such as CLAWS (Constituent Likelihood Automatic Word-tagging System), which is free to use for corpora up to 100,000 words. I entered the example sentence below into the text box at http://ucrel.lancs.ac.uk/claws/trial.html and the tagged sentence, with PoS (part of speech) codes after each word, looked like this:

Here_AV0 is_VBZ a_AT0 simple_AJ0 example_NN1 of_SENT tagging_VVG

**Corpus Search**

- Go to: http://ucrel.lancs.ac.uk/claws/trial.html
- Enter a short text of your choice in plain text format.
- Congratulate yourself on having tagged a text (if you haven’t done so before).

The tagging code can be accessed at http://ucrel.lancs.ac.uk/claws5tags.html. It is important to note that the tagging system goes beyond the basic parts of speech: there are over 150 tags in some versions of CLAWS. A tagged corpus facilitates certain kinds of searches: you might, for example, want to search for all the examples of set used as a noun.

**Analysing corpus data**

Once you have saved your principled collection of texts, you have a corpus. Now you need the software to analyse it. Such software can be downloaded free from the internet (often with licence restrictions) – antconc (http://www.antlab.sci.waseda.ac.jp/software.html) is an example of this type of software. Wordsmith Tools (http://www.lexically.net/wordsmith/) has established itself as an industry standard and is available for purchase. The three basic analytical operations are word frequency counts, concordance and collocation. Frequency lists are easy to generate. The examples of analysis we look at below were produced through the free-to-access interface to COCA and BNC at http://corpus.byu.edu/bnc/

**Word frequency counts**

Below are frequency lists from COCA (http://www.wordfrequency.info/free.asp?s=y). Note that there is both a rank order list and an alphabetical list (the alphabetical list also has rank order information).
Building a corpus

**Figure 2.1 COCA word frequency list**

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Alphabetical</th>
<th>Part of Speech (PoS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rank</td>
<td>word/lemma</td>
<td>PoS</td>
</tr>
<tr>
<td>1</td>
<td>the</td>
<td>a</td>
</tr>
<tr>
<td>2</td>
<td>be</td>
<td>v</td>
</tr>
<tr>
<td>3</td>
<td>and</td>
<td>c</td>
</tr>
<tr>
<td>4</td>
<td>of</td>
<td>i</td>
</tr>
<tr>
<td>5</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>6</td>
<td>in</td>
<td>i</td>
</tr>
<tr>
<td>7</td>
<td>to</td>
<td>t</td>
</tr>
<tr>
<td>8</td>
<td>have</td>
<td>v</td>
</tr>
<tr>
<td>9</td>
<td>to</td>
<td>i</td>
</tr>
<tr>
<td>10</td>
<td>it</td>
<td>p</td>
</tr>
<tr>
<td>11</td>
<td>l</td>
<td>p</td>
</tr>
<tr>
<td>12</td>
<td>that</td>
<td>c</td>
</tr>
<tr>
<td>13</td>
<td>for</td>
<td>i</td>
</tr>
<tr>
<td>14</td>
<td>you</td>
<td>p</td>
</tr>
<tr>
<td>15</td>
<td>he</td>
<td>p</td>
</tr>
<tr>
<td>16</td>
<td>with</td>
<td>i</td>
</tr>
<tr>
<td>17</td>
<td>on</td>
<td>i</td>
</tr>
<tr>
<td>18</td>
<td>do</td>
<td>v</td>
</tr>
<tr>
<td>19</td>
<td>say</td>
<td>v</td>
</tr>
<tr>
<td>20</td>
<td>this</td>
<td>d</td>
</tr>
<tr>
<td>21</td>
<td>they</td>
<td>p</td>
</tr>
<tr>
<td>22</td>
<td>at</td>
<td>i</td>
</tr>
<tr>
<td>23</td>
<td>but</td>
<td>c</td>
</tr>
<tr>
<td>24</td>
<td>we</td>
<td>p</td>
</tr>
<tr>
<td>25</td>
<td>his</td>
<td>a</td>
</tr>
</tbody>
</table>

**Concordancing**

The concordancing tool displays all the instances of the word or phrase you are looking for in the corpus with a limited amount of co-text either side of the target word. It is normal to display the target word or phrase in the middle of the concordance line, often highlighted. Below is an example of a concordance search for the word *dulcet* from the COCA spoken component accessed at http://corpus2.byu.edu/coca/

Note how the search term is entered on the left-hand part of the screen and the component of the corpus specified just below.
Here we can see that *dulcet* has a tendency to collocate with the word ‘tone’ in spoken language. Note that the words to the right of the node word *dulcet* are arranged alphabetically. It is possible to switch this such that the words to the left are arranged alphabetically. As O’Keeffe, McCarthy and Carter (2007) note, when reading concordance output you need to get into the habit of reading from the middle outwards rather than from left to right.

**Collocations**

From the concordance lines, we have a hypothesis that *dulcet* collocates frequently with *tones*. We can check this hypothesis by carrying out a collocation search on the whole corpus, as illustrated below. Note that the search is carried out by entering *dulcet* in the word box, clicking the question mark next to *dulcet* and...
Figure 2.4 BNC collocation search: *dulcet*
determining the number of words to the left and right of *dulcet* where the search should be carried out, i.e. in this case we are interested in the most frequent collocates which occur within four words to the left or four words to the right of our target word *dulcet*. Note that in this case we search the whole corpus, not just the spoken component.

We can see immediately that our hypothesis is confirmed, i.e. *tones* is the most common collocate of *dulcet*. We might then want to investigate whether this is the case in British English and so on.

We could also make a more general observation based on the collocation frequency list above: *dulcet* has a tendency to collocate with words which describe sounds, e.g. *drawl*, *moan*, i.e. there is a semantic link between many of the collocates of *dulcet*.

**Corpus Search**

If you would like to familiarise yourself a little more with basic corpus searches, you could repeat the same three searches above (frequency, collocation and concordance) using the BNC at the same site (rather than COCA). Go to: http://corpus.byu.edu/bnc

**References**


Chapter 3

Corpora and lexis

Introduction

The focus of this chapter is on the contribution corpus research has made both to our understanding of the nature of the lexicon and to our knowledge of the behaviour of words and other kinds of lexical item which together comprise the lexicon. The value of corpora for lexical research is underlined by Krishnamurthy (2000) who points to the limitations of linguistic descriptions which are based only on our intuitions about language; he observes in particular that intuition can lead to imprecise and inaccurate judgements about lexis. More specifically, Krishnamurthy (2000) argues that an approach to description based only on intuition, can lead to a tendency to ‘notice unusual words or structures but often overlook ordinary ones’. Corpus research, as we shall see, suggests that it would be wise for ELT researchers and practitioners to focus on the meaning and behaviour of ‘ordinary words’, perhaps more so than we have done in the past.

It is important to note at the outset that corpus research has not only led to more detailed descriptions of the meaning and behaviour of words; it has also suggested that we need to consider units beyond the word in describing the lexicon. Jones and Durrant (2010: 389), for example, point to the role of corpus linguistics in identifying units beyond the word:

One of the central insights to come from corpus linguistics in the last thirty years is the extent to which competent users draw not only on a lexicon of individual words, but also on a range of lexicalised phrasal units which have come to be known as ‘formulaic sequences’.

Such insights, Greaves and Warren (2010: 221) argue, have led to nothing less than a ‘reappraisal of the status of lexis’ which entails a reconceptualisation of both the relative roles of lexis and grammar in communication and of the relationship between lexis and grammar. This reappraisal of the status of lexis is quite radical. While grammar has traditionally been seen as the generative aspect of language and, therefore, prioritised in English language teaching, Sinclair et al.
Corpora and lexis 23

(2004: xxv), for example, see grammar as playing only a subsidiary or supportive role in communication:

… on the whole grammar is not involved in the creation of meaning, but rather concerned with the management of meaning.

The shift of emphasis from grammar to lexis in accounting for language production is neatly summed up by Wilkins (1972: 111):

Without grammar little can be conveyed; without vocabulary, nothing can be conveyed.

Writing in the context of ELT, Lewis (1993: iv) also argued that grammar is subordinate to lexis in creating meaning, a view captured in his oft-quoted aphorism:

Language consists of grammaticalised lexis, not lexicalised grammar.

This principle that lexis is primary in communication can be illustrated very simply. Below, I have removed the lexical words from the opening sentence to this chapter. It is impossible to glean any sense from it (unless you have an extraordinary memory!).

Has; both; to; our; of; the; of; the; and; to; our; of; the; of; and; which; the

Faced with just the lexical items from the same sentence, however, as below, we can at least begin to reconstruct the meaning of the sentence.

Corpus; research; contribute; significant; understanding; lexicon; knowledge; behaviour; words other lexical items comprise lexicon

With regard to the relationship between lexis and grammar, Römer (2009) speaks of the ‘interdependence of lexis and grammar’, domains which have traditionally been regarded as quite separate, with the result that, in ELT, grammar has often been taught without much thought to the lexis used to illustrate and practise a given structure, while lexis has been taught without much thought to how it operates grammatically. We will, however, postpone discussion of the interdependence of lexis and grammar until Chapter 4.

In this chapter, we consider quantitative and qualitative corpus-informed reappraisals of the nature and status of lexis and their implications for ELT, focusing on notions such as collocation, lexical chunks and semantic preference. Throughout the chapter we consider the implications of lexically-oriented research for ELT practice.
Describing the lexicon

Terminology

In our discussion of the lexicon we need to be clear about the terms we are using: indeed, even the notion of a ‘word’ needs some clarification. It is customary to distinguish between a word (form) and a lemma. A lemma is the basic form of the word: the lemma ‘meet’ for example has the associated word forms meets, met and meeting. This distinction between ‘word’ and ‘lemma’ is relatively unproblematic. Unfortunately, that is not the case with multi-word items, where we meet a wide variety of terms which often overlap without being coterminous. We have already encountered four in this introduction: ‘lexicalised phrasal units’, ‘formulaic sequences’, ‘collocations’ and ‘lexical chunks’. Other common terms to describe multi-word units of one kind or another in the literature include ‘colligation’, ‘lexical phrase’, ‘lexical chunk’, ‘lexical bundle’, ‘formulaic sequence’ and ‘prefabricated unit’. Given that our concerns are pedagogic rather than purely descriptive, it makes sense, I would argue, to limit the number of terms on the lines suggested below with the proviso that in referring to the work of specific authors, we will need to quote the terms they choose to use in their work (see also Timmis 2008). I propose to operate with the definitions below:

1 Collocation: a combination of two lexical (as opposed to grammatical) words often found together or in close proximity, e.g. dire straits; make sense. We need to note three other points in relation to this definition:

a If the pair of words is separated by an article as in have a party, this still counts as a collocation in our definition.

b Phrasal verbs such as carry out can be treated as one word, so carry out an experiment still counts as a collocation in our definition.

c Some researchers refer to the kind of collocation we have defined and exemplified above specifically as a lexical collocation.

Corpus Search

1 List what you think are the five most common noun collocates of ‘carry out’ on the right hand side, e.g. carry out a task (task is the right-hand collocate of carry out in this case).

2 Now go to http://corpus.byu.edu/bnc/ to check your intuition.

3 If you click the question mark by collocates, you will find this display:
4 Enter ‘carry out’ in the **word box**.
5 Enter 0 in the left hand box by **collocates** but leave 4 in the right hand box.
6 Click on **POS** and select noun (all).
7 You are now set to search for the most frequent collocates to be found in the corpus within four words to the right of the word **carry out**.

8. Click **search**.
The next item we need to define is a lexical chunk, which I am going to define as a frequent meaningful sequence of words that may include both lexical and grammatical words, e.g. ‘to a certain extent’ includes a preposition and an article.

Lexical chunks, as Schmitt (2000) notes, include features such as phrasal verbs, idioms and compounds which have traditionally been recognised as multi-word units. However, as the examples below show, the term ‘lexical chunk’ may cover many other sequences of words commonly found together, such as:

- to a certain extent
- by the way
- last but not least
- at the end of the day
- play a role
- not only … but also
- depend on.

We should note here that some chunks are invariable, e.g. ‘by the way’; others are minimally variable, e.g. ‘to a certain extent/degree’; others are more freely variable with an open slot, e.g. play a/an + adjective + role. We should also note that our definition of lexical chunk effectively subsumes other terms one finds in the literature such as lexical phrase, phrasal expression, formulaic sequence and lexicalised sentence stem.

Lexical bundle: a sequence of words found together without a clear semantic or pragmatic meaning.

To take an example of a lexical bundle which meets our definition, Biber et al. (1999), note that it was in the is a commonly recurring string of four words, but this would not count as a collocation or a lexical chunk in our definition as it does not hold much self-contained meaning. Shin and Nation (2008), for example, note that ‘The five most frequent 2-word groups in the Brown Corpus are of the; in the; to the; on the; and the’. These 2-word combinations do not meet Shin and Nation’s meaningfulness criterion for collocation, nor do they meet our criteria: they are ‘lexical bundles’ in our terms (though I have to concede that some researchers do not distinguish between meaningful and meaningless sequences in their definition of lexical bundle).

Given that our definitions have been accompanied by discussion, it will be useful to repeat them here for clarity and ease of reference:

1. **Collocation**: a combination of two lexical (as opposed to grammatical) words often found together or in close proximity, e.g. dire straits; make sense.
2 **Lexical chunk**: a frequent meaningful sequence of words which may include both lexical and grammatical words, e.g. ‘to a certain extent’ includes a preposition and an article.

3 **Lexical bundle**: a sequence of words found together without a clear semantic or pragmatic meaning.

**Discussion**

How important is it that learners understand such terminology?

The terminology we have discussed above is, I would argue, the most useful for pedagogically-oriented corpus work and these are the terms we will use most often in this book. You may then wish to go to section 2.2 now. However, I have no wish to limit the ambitions of readers, so two further descriptive terms for lexical items are briefly discussed below:

**Ngrams** are word strings of any specified length. You can search, for example, for 2-word, 3-word, 4-word sequences and so on. The list produced will include both what we term ‘lexical bundles’ and what we term ‘lexical chunks’. I ran an ngram search using this chapter as my corpus. I specified ngrams between three and five words long using the concordance programme ‘antconc’ (http://www.laurenceanthony.net/software.html). The results are shown in Figure 3.3 and we can see that it includes both lexical chunks, e.g. ‘I would argue’ and meaningless bundles such as ‘O’Keeffe et’. We can note in passing that the list seems to show that I am an unusually helpful person.

**concgram** identifies all of the co-occurrences of specified words within a wide span. Unlike ngrams, the words do not need to be contiguous: ‘not all of these instances are necessarily meaningfully associated’ (Greaves and Warren 2010: 218). A concgram search, for example, might show that, in a given corpus, *I think* and *because* are often found within a reasonably wide span of each other, e.g. ten words or fewer – Greaves and Warren (2010) argue that concgram can be useful in identifying typical ‘organisational frameworks’ in certain kinds of writing. See http://www.lexically.net/downloads/version5/HTML/index.html?definition_of_a_concgram.htm

The example above of *I think* and *because* as concgrams is a suitable point to observe that corpus research is not necessarily pedagogically useful. The fact that thoughts are often accompanied by reasons is hardly surprising in a world
where rational thought is respected, and the *I think – because* frame is just about the most obvious way of doing it (indeed, to be fair, this may be the very reason Greaves and Warren [2010] chose it as an introductory example). This is not to denigrate concgram research per se, but a reminder that we need to look at all corpus research through a pedagogically oriented critical filter and ask not just, ‘What can a corpus do?’ but, ‘What can a corpus do for us?’
Applying our terminology

Collocations in context

The definition of ‘collocation’ above – two lexical words often found together or in close proximity to each other – is applied below to a paragraph from the introduction to this book, highlighting some of the examples which meet our definition. The collocations highlighted are ones I identified intuitively before checking at http://corpus.byu.edu/bnc/

This comment by Conrad (2000) encapsulates the main aim of this book: to help move corpus linguistics from what Römer (2012) terms its ‘minority sport’ status in language teaching to a point where the ability to carry out and interpret corpus research is seen as a normal part of an English language teacher’s repertoire. Familiarity with corpus research and practice should be a standard part of an English language teacher’s toolkit, I would argue, because most people in ELT will at some time have had thoughts like these:

• Is this grammar rule in the coursebook really right?
• Do my students really need this grammar point?
• What vocabulary do my English for dentistry students need to get their teeth into?
• Why is everyone talking about lexical chunks and collocations?

If you have had questions like these, this book is designed to help you to answer them. It is also designed to help you to generate and investigate similar questions.

The book does not assume prior knowledge or experience of corpus research; nor does it assume any technical expertise. Technophobes can relax: contemporary corpus interfaces and corpus software are user-friendly and often include tutorial packages. If you know how to save a document, you are, as we shall see, well on the way to being able to compile a corpus; and then things get really interesting.

We can make a number of observations based on this text analysis of collocations:

1 We find collocations made up of different grammatical combinations
   • Adjective + noun collocations, e.g. main aim; technical expertise; prior experience; prior knowledge
   • Noun + noun collocations, e.g. grammar rule
   • Verb + noun collocations, e.g. save a document; compile a corpus

2 Collocates are not necessarily immediately adjacent, e.g.
   • If you have had questions like these, this book is designed to help you to answer them;
   • … to a point where the ability to carry out and interpret corpus research is seen as a normal part of …
You have to decide which word you are treating as the ‘node word’. With main aim for example, you could look for nouns to the right-hand side of main (and you find aim in first place); alternatively, you could search for adjectives to the left-hand side of aim (and you find main in 14th place).

As we have defined a collocation as consisting of two lexical words, it is open to question whether have a question or have a thought are collocations or lexical chunks as have can be regarded as a ‘delexical’ verb, i.e. a verb with low lexical content whose meaning is conditioned in context by the words it co-occurs with (O’Keeffe, McCarthy and Carter 2007).

Some collocations are field-specific, e.g. compile a corpus; grammar rule.

The selection of collocations above is indicative rather than exhaustive.

The phrasal verb carry out is counted as one word.

There is no suggestion that these collocations are of equal teaching/learning value. At this stage we are only concerned with illustrating the concept.

---

**Corpus Search**

1. Take a paragraph from a text you have in electronic form.
2. Identify four or five collocations intuitively.
3. Go to http://corpus.byu.edu/bnc/
4. Check your intuition. Remember to decide which part of the collocation you are treating as the node word. The node word is the word which attracts the collocate, e.g. in our earlier example carry out was the node word and task the collocate. Remember to specify the part of speech you are looking for (by clicking on the PoS icon).

```
Figure 3.4  BNC search: carry out + noun
```
Lexical chunks in context

Our definition of ‘lexical chunk’ above is applied below to the same paragraph from the introduction to this book, highlighting some of the examples which meet our definition. As above, these chunks were initially identified intuitively and then checked at http://corpus.byu.edu/bnc/

This comment by Conrad (2000) encapsulates the main aim of this book: to help move corpus linguistics from what Römer (2012) terms its ‘minority sport’ status in language teaching to a point where the ability to carry out and interpret corpus research is seen as a (normal) part of an English language teacher’s repertoire. Familiarity with corpus research and practice should be a (standard) part of an English language teacher’s toolkit, I would argue, because most people in ELT will at some time have had thoughts like these:

• Is this grammar rule in the coursebook really right?
• Do my students really need this grammar point?
• What vocabulary do my English for dentistry students need to get their teeth into?
• Why is everyone talking about lexical chunks and collocations?

If you have had questions like these, this book is designed to help you to answer them. It is also designed to help you to generate and investigate similar questions.

The book does not assume prior knowledge or experience of corpus research; nor does it assume any technical expertise. Technophobes can relax: contemporary corpus interfaces and corpus software are user-friendly and often include tutorial packages. If you know how to save a document, you are, as we shall see, (well) on the way to being able to compile a corpus; and then things get really interesting.

1 A collocation can ‘nest’ within a chunk, e.g. the main aim of; this could be further extended to a lexicalised sentence stem, i.e. the main aim of x is to y.
2 A lexical chunk can have an ‘optional slot’, e.g. a/an + adjective + part + of (a normal/standard part of).
3 Our notion of lexical chunks includes sentence frames such as ‘move x from y to z’, e.g. move (1) corpus linguistics from (2) what Römer (2012) terms its ‘minority sport’ status in language teaching to (3) a point where.
4 Our notion of lexical chunks includes idioms such as get their teeth into.
5 When teaching it can be useful to record grammatical information with variable lexical chunks, e.g. on the way to + gerund/noun (on the way to being able to compile a corpus).
6 This selection of chunks from the text is indicative rather than exhaustive.
7 The chunks highlighted will not be of equal teaching/learning value. As was the case with collocations, at this stage we are only concerned with illustrating the concept.
Chunks can be domain-specific, e.g. *I would argue* is common in academic English.

### Corpus Search

1. Take a paragraph from a text you have in electronic form (the same one you chose for the previous task will be fine).
2. Identify three or four chunks intuitively.
3. Go to http://corpus.byu.edu/bnc/
4. Enter your chunk in the word box (no need to click on anything else).
5. Click submit.
6. See how many ‘hits’ you get for your chosen chunk.

Hopefully, the terms *collocation* and *lexical chunk* will be serviceable for pedagogic purposes, facilitating both our discussions in this chapter and, just as importantly, discussion with learners. However, precise use of terminology is not an absolute priority with learners, as Conzett (2000: 85) argues:

> The single most important thing for teachers, more than worrying whether or not something is a collocation, is to shift their and their students’ focus away from individual words to chunks of language.

In the text below, the selection of collocations and lexical chunks that were highlighted separately above are brought together to illustrate the pervasiveness of collocations and lexical chunks in our text. Collocations are highlighted in **bold**; lexical chunks are highlighted in **bold italic**.

This comment by Conrad (2000) encapsulates the main aim of this book: to help move (1) corpus linguistics from (2) what Römer (2012) terms its ‘minority sport’ status in language teaching to (3) a point where the ability to carry out and interpret corpus research is seen as a (normal) part of an English language teacher’s repertoire. *Familiarity with* corpus research and practice should be a (standard) part of an English language teacher’s toolkit. *I would argue*, because most people in ELT will **at some time** have had thoughts like these:

- Is this grammar rule in the coursebook really right?
- Do my students really need this grammar point?
- What vocabulary do my English for dentistry students need to **get their teeth into**?
- Why is everyone **talking about** lexical chunks and collocations?
If you have had questions like these, this book is designed to help you to answer them. It is also designed to help you to generate and investigate similar questions.

The book does not assume prior knowledge or experience of corpus research; nor does it assume any technical expertise. Technophobes can relax: contemporary corpus interfaces and corpus software are user-friendly and often include tutorial packages. If you know how to save a document, you are, as we shall see, (well) on the way to being able to compile a corpus; and then things get really interesting.

For the sake of completeness, I reproduce the result of a n-gram search on the same text in Figure 3.5. I set the n-gram search to look for ngrams of between two and six words as a variation from the example in the introduction where the search was for ngrams of between three and five words.

Figure 3.5 N-gram search 2 using antconc software
Corpus Question

Operating our definitions of lexical chunks and lexical bundles in this chapter, which of the ngrams in the screenshot would you define as collocations, which as lexical chunks and which as lexical bundles?

Discussion

How far do you think ngram corpus searches have pedagogic potential?

The importance of lexis – corpus perspectives

Lexis and fluency

In the previous section we considered how different types of lexical unit might be defined, settling on four basic terms: word, collocation, lexical chunk and lexical bundle. In this section we consider how such descriptive work, with its insights into word frequency and the tendency of words to co-occur in collocations, lexical chunks and other multi-word items, has led to changing views of the principles of language production and reception. While we argued earlier that specific corpus-based findings may or may not be relevant to ELT (e.g. the co-occurrence of I think and because), and that corpora which represent a specific variety of English may not be relevant to all learners, research that throws light on fundamental issues such as how language is produced, received and understood would appear to be of obvious interest to anyone engaged in ELT.

A key concept to consider in relation to the more general significance of corpus-based lexical research is Sinclair’s (1991) distinction between the ‘idiom principle’ and the ‘open choice principle’ as principles of language production. The basis of the idiom principle, as defined by Sinclair (1991: 110) is that ‘a language user has available to him or her a large number of semi-preconstructed phrases [i.e. collocations and chunks in our terms] that constitute single choices, even though they might appear to be analysable into segments’. The open choice principle (or ‘slot and filler’ approach), on the other hand, involves direct recourse to grammar and word-by-word generation of an utterance or sentence. The significance of this distinction for our purposes is that it has been argued (e.g. Pawley and Syder 1983, Nattinger and deCarrico 1992, Lewis 1993) that the idiom principle underpins fluent language use. Fluency is gained by making fewer and larger choices from the lexicon than one would make by following the ‘slot and filler’ approach, as the open choice principle is sometimes known. The role of lexical chunks in achieving fluency is stressed by Nattinger and deCarrico (1992: 32):
It is our ability to use lexical phrases that helps us to speak with fluency. This prefabricated speech has both the advantages of more efficient retrieval and of permitting speakers (and learners) to direct their attention to the larger structure of the discourse, rather than keeping it narrowly focused on individual words as they are produced.

While Nattinger and deCarrico point to the importance of collocations and chunks for production, we also need to keep in mind Willis’ (2003: 44) argument that a stock of chunks is important for both productive and receptive purposes:

Efficient communication is, then, not simply a matter of making any grammatical sentence. It depends on having a stock of fixed phrases which we can string together rapidly and efficiently … It is also a matter of recognising and producing familiar forms of speech which can be readily processed … (my italics)

Let us look at an example of how we call on a stock of chunks to produce utterances: if you ask any native speaker to give you an example of the word ‘figment’ in use, they will almost certainly produce an example like this (I know because I have tried it with a number of native speakers):

- It is/was just/only a figment of his/your etc. imagination.

The fact that I can predict that native speakers will produce this phrase with the word ‘figment’ makes it highly likely that it is stored and accessed as a whole phrase (with slight variations) rather than strung together word by word on a given occasion. This is also relevant to our point about the role of collocations and chunks in reception. As soon as I hear or read the word figment I can predict the rest of the phrase with a high degree of accuracy. The phrase, ‘It is merely a product of your dreams’, may express the same meaning would be understood, but it will not be processed as quickly as it is simply not conventional. Similarly, to borrow an example from Willis, if I ask you the time and you say, ‘It’s half before two’, it will take me fractionally longer to process this than if you say, ‘it’s half past one’ as the latter form is the convention in British English – it meets my expectations. There are times, however, when we may deliberately choose to break conventions for specific effect, but these will be the exception rather than the rule.

As Peters (1983) argues, it is likely that, in addition to a grammar and a lexicon (vocabulary store) in the brain, we have (metaphorically speaking) ‘a phrase book with grammar notes’. In these terms, the ‘figment chunk’ we proposed above would be stored in our ‘brain phrase book’ something like this:

- It (or full noun phrase) + to be + (adverb – just/simply/only) a + figment + of + possessive + (adjective) + imagination.
What this analysis suggests in practical classroom terms, I would argue, is that if a learner asks you, ‘What does ‘figment’ mean?’ there is almost no point in answering the question unless you give examples of the chunk(s) in which it is typically used. Indeed, when I have asked native speakers what the word *figment* means, none has been able to answer immediately, which suggests that they know it only as part of a lexical chunk with holistic meaning. *Figment* is perhaps a marked example of a word which seems only to derive its meaning from context, but we should note Moon’s (2010: 202) argument that, ‘By forcing us to consider words in context, corpus concordances make us aware of how far the meanings of words are derived from context – even raising the question of whether words have independent meaning at all’.

**Corpus Search**

1. Run a corpus search on the word ‘figment’ at http://corpus.byu.edu/bnc/
2. From the concordance lines choose two or three examples which you think would be good to illustrate the meaning for intermediate students.
3. Explain your criteria for your choice of examples.

The importance of collocations and chunks for both reception and production suggests that acquiring a stock of such items will be of great value for English language learners. As O’Keeffe, McCarthy and Carter (2007: 60) point out: ‘… for the learner of any second or foreign language, learning the collocations of that language is not a luxury if anything above a survival mastery of the language is desired’.

**Layers of meaning**

Thus far we have considered the contribution of corpus research to the identification of lexical units such as collocations and lexical phrases and identified the important role such units play in fluent communication. We need now to consider corpus research that has shed new light on the kinds of meaning words and phrases can carry. Key notions in this respect are ‘semantic preference’, ‘semantic prosody’ and ‘lexical priming’, which we discuss in turn below.

‘Semantic preference’ refers to the tendency of a word to collocate with sets of semantically related collocates (Partington 2004 in McEnery *et al.* 2006: 148). Partington (2004) exemplifies this tendency through the semantic preferences of the adjective *sheer*, illustrating that the collocates of *sheer* tend to fall into four semantic groups, as exemplified below.
1. volume/weight, e.g. sheer magnitude
2. force/energy, e.g. sheer strength
3. persistence, e.g. sheer determination (here structures such as ‘through sheer determination’ are common)
4. strong emotion – sheer joy.

### Corpus Question

Apply Partington’s categories above to the concordance printout below:

**Figure 3.6 BNC KWIC concordance: sheer**

From a teaching point of view, it is also worth considering Moon’s (2010: 205) notion of ‘canonical semantic structures’. In example a) below, we see that the noun *refuge* typically collocates with the verb *take*, forms a chunk with the preposition *from* and is semantically associated with ‘undesirable situations’ and ‘safe places’. In example b) we can see that the verb *comply* typically occurs with a human subject, is often preceded by a modal verb, forms a chunk with the preposition *with* and semantically typically co-occurs with ‘official instructions’.

```plaintext
a  (take) refuge from + undesirable situation + in + safe place
b  Someone (+ modal) + complies with + official instruction
```
Corpus Question

Look at the list of collocates below for the verb come up with. What does the list of collocates tell us about the semantic preference of come up with?

Figure 3.7 BNC collocation search: come up with
(See Appendix 1, Commentary 2)

‘Semantic prosody’ refers to the meaning a node word (the word which attracts the collocates) takes on because of the presence of a consistent set of collocates (Louw 2005). The example cited by Partington (2004) is that happen tends to co-occur with unpleasant events, as illustrated below through a BNC search for the noun collocates of happen to the left-hand side (i.e. nouns which typically precede happen).

Figure 3.8 BNC collocation search: happen

From the left-hand part of the screen, we can see that the search was set to look for noun collocates within four words (on the left-hand side) of the node word ‘happen’. We can immediately see when we look at the list of collocates on the
right-hand part of the screen that ‘accident(s)’ is a very frequent collocate of \textit{happen}. But we need to make two further observations:

1. \textit{Thing} and \textit{things} are the most frequent collocates of \textit{happen}, but to determine the semantic prosody, we would need to examine the concordance lines with ‘thing(s)’ qualitatively to determine in each instance whether \textit{thing} denotes a positive or negative phenomenon.

2. While the common collocates \textit{accident} and \textit{accidents} suggest that \textit{happen} has a negative semantic prosody, the fact that \textit{miracle} – generally a positive word – is also a collocate of \textit{happen}, underlines that semantic prosody is a generalised tendency rather than an absolute rule: we \textit{cannot} say that \textit{happen} always collocates with negative nouns. As is so often the case with corpus data, we have to talk in probabilistic terms.

As a further example of semantic prosody, Ruehlemann (2010) refers to the tendency of the verb ‘set in’ (in its meaning ‘to become fixed’) to be used with negative situations: a BNC search reveals that ‘panic’, ‘decline’, and ‘rot’ are frequent left-hand collocates of ‘set in’. Indeed, the prototypical example of \textit{set in} seems to be ‘The rot has set in’, based on a quick survey when I asked colleagues to give me an example with ‘set in’.

We noted in relation to collocations and chunks above that the very fact that they are routinised forms offers creative and rhetorical potential in breaking them. The same applies in relation to semantic prosody: Ruehlemann (2010), for example, notes the tendency of ‘break out’ to occur with nouns with a meaning of physical violence, e.g. ‘War has broken out’. The BNC search below ([nn* broke out] clearly shows that \textit{break out} typically has a negative semantic prosody.

\textbf{Figure 3.9} BNC collocation search: \textit{broke out}
Precisely because of this expectation, deviations from semantic prosody can be used to create ironic effect, e.g. in the BNC extract below (from the *Belfast Telegraph* newspaper), there is a clear implication that the province did not previously enjoy ‘an excellent industrial relations record’.

*Peace has broken out* in relations between workers and employers in Northern Ireland, figures revealed today. Speaking at the opening of new offices for the Industrial Tribunals and Fair Employment Tribunals in Belfast, Economy Minister Robert Atkins said the province now enjoyed an excellent industrial relations record.

**Corpus Question**

The screenshot below lists the right-hand noun collocates of a specific verb. Can you guess what the verb is which can precede all these nouns? What do we learn about the semantic prosody of the word in question?

![Figure 3.10 BNC collocation search: mystery word](See Appendix 1, Commentary 3)

**Discussion**

How useful do you think the notions of semantic preference and semantic prosody exemplified above are for ELT practitioners? How far do you think semantic preference, semantic prosody and semantic structure are teachable?
We have seen that corpus research shows that our knowledge of a word extends beyond its basic meaning and can include knowledge of its common collocates, of semantic relations between these collocates, and a feeling for the covert meanings with which a word is imbued by its typical collocates (semantic prosody). Hoey’s (2005: 8) notion of lexical priming builds on this complex picture of word knowledge, arguing that our knowledge of a word builds incrementally into a rich web of information:

A word is acquired by encounters with it in speech and writing. A word becomes cumulatively loaded with the contexts and co-texts in which they are encountered. Our knowledge of the word includes the fact that it co-occurs with certain other words in certain kinds of context. The same process applies to word sequences built out of those words; these too become loaded with the contexts and co-texts in which they occur.

Lexical priming suggests, then, that we can add to aspect of word knowledge we have already identified – meaning, collocational behaviour, semantic preference and semantic prosody – other kinds of word knowledge: the language structures, textual positions, or text types the word typically appears in (Römer 2009).

We have established in this section that corpus linguistics has presented a different picture of the lexicon in which it is possible to identify a number of lexical units beyond the word. We have also established that this picture of the lexicon has potential implications for language teaching as fluent production (and reception) is likely to be facilitated by the ability to access collocations and chunks efficiently. Finally, we have noted that words can carry layers of meaning such as semantic prosody which are not necessarily intuitively obvious and which add another dimension to the lexical challenges learners face. In the next section we turn from general reflections on the nature of the lexicon to specific quantitative insights of potential value to ELT.

**Corpora and frequency**

Perhaps the most obvious application of a corpus in the field of lexis is its ability to produce automatically more comprehensive and reliable frequency lists than can be generated manually. As O’Keeffe, McCarthy and Carter (2007: 31) point out, from corpus frequency information, we can ‘ascertain how many words native speakers use, how frequently they have recourse to the individual words they use and how they combine them, and to explore to what extent words have become part of regularly occurring chunks or clusters for the native user’. The same point about the kind of information we can obtain also applies, of course, not only to corpora representing native speaker language use, but also to corpora representing learner or expert user language.

In this section, we review corpus frequency findings in relation to words, collocations and lexical chunks, discussing the pedagogical implications of these
frequency findings for ELT in turn. Although, for convenience, we discuss frequency lists for words, collocations and chunks separately in this section, we will discuss later the case for integrating these lists.

**Word frequency**

One of the most striking corpus findings in terms of word frequency with ramifications for ELT is that ‘a relatively small number of very common items accounts for the large majority of language we typically encounter’ (Jones and Durrant 2010). Nation (2013), for example, notes that the 2,000 most frequent words in English account for 80–90% of language in conversations or books. Word frequency findings seem to offer us, then, a starting point for the construction of a vocabulary syllabus. The most frequent 2,000 words of the language, for example, is generally seen as a good threshold target for learners as these words constitute around 80% of texts on average (e.g. Nation 2013). Moving up the proficiency levels, O’Keeffe, McCarthy and Carter (2007: 48–49) suggest that ‘… a receptive vocabulary of some 5-6,000 words would appear to be a good threshold at which to consider learners to be at the top of the intermediate level …’, as this figure constitutes around 90% of texts on average. At advanced levels, O’Keeffe, McCarthy and Carter (2007) suggest that it may be more a question of learning more about the words they already know than learning ever more obscure new words of doubtful utility: this may involve learning less frequent or metaphorical meanings of words they are already familiar with. This suggestion becomes even more pertinent when we consider Moon’s (2010: 201) observation:

... much of the corpus evidence for high-frequency verbs such as *take* and *get* consists of occurrences in phrasal verbs and other fixed expressions, while an important feature of data for nouns such as *hand*, *head* and *heart* is the way in which they are used metonymically and metaphorically, including in idioms.

The words *take*, *get*, *hand*, *head* and *heart* would almost certainly feature in an elementary vocabulary syllabus, but, as Moon (2010) suggests, there is a lot to learn about how they are used in combination with other words. Knowing the word *foot*, for example, does not lead to automatic comprehension of the idiom *to foot the bill*. As O’Keeffe, McCarthy and Carter (2007: 54) note, we need to consider breadth of knowledge as well as depth of knowledge: ‘What needs to happen alongside the increase in breadth is an increase in depth of knowledge, i.e. the knowledge of the various aspects of use of a word, including, beyond its formal properties, its collocations, its sub-senses, and its semantic prosody’.
Discussion

In view of this comment by O’Keeffe, McCarthy and Carter (2007) about breadth and depth of knowledge, what do learners need to know about the verb arise before they can be said to really know it? (see Appendix 1, Commentary 4)

Corpus Search

Carry out a corpus search of arise at http://corpus.byu.edu/bnc/ and see if there are any further aspects of this word you might draw attention to. (see Appendix 1, Commentary 4)

While the main aim of this book is to illustrate the potential of corpus linguistics to contribute to ELT, we should also note its limitations in order to see this potential in perspective. First, the value of a word frequency list to a learner is relative to the corpus it is based on: a word frequency list from a general British English corpus such as the BNC, for example, will be of limited value to a learner who wants to study engineering in an English medium university in India. We keep in mind here McCarten’s (2007: 4) observation that word frequency has limitations as a vocabulary selection criterion (and this will apply to any corpus we may choose to consider).

It may not be possible to use all the items in the list for a number of reasons. Some may be culturally inappropriate, not suitable for class, or just difficult to use until students have more English. Also, the communication needs of students may be different from those of the people whose conversations are recorded in the corpus. For example, a word like homework, a frequent word in any classroom, comes toward the end of the top 2,000 words, whereas words like supposed, true, and already, which are in the top 400, might be challenging for elementary learners.

Another clear reason why word frequency cannot be the only criterion is pedagogic convenience: it would be absurd, for example, to separate days of the week in the vocabulary syllabus because some days are more frequently mentioned than others in the corpus (McCarten and McCarthy 2010). Similarly, many of the high frequency words in a corpus are grammatical function words rather than the lexical content words learners need to convey basic meaning (Nation 2013). It is important, then, to bear in mind Jones and Durrant’s (2010) observation that important decisions need to be made before we can make an effective link between corpus findings and the vocabulary syllabus. Specifically, Jones and Durrant (2010) refer to three important decisions we need to make in relation to pedagogic applications of word frequency counts:
Does the corpus we have chosen or constructed match the needs of our learners? The word frequency of a British English corpus, for example, will be determined in part by aspects of British culture, which might be quite irrelevant to the context and needs of a specific group of learners.

Are we going to base our frequency list on individual word forms or lemmas, i.e. do we treat book, books and booked as separate words or instances of one lemma – book?

How are we going to deal with polysemous words such as set? Do we try to distinguish the frequencies of each separate meaning?

Thus far we have discussed word frequency findings mainly in relation to the vocabulary syllabus, but it is not just at the level of syllabus design that corpora have the potential to contribute to vocabulary teaching; corpora have allowed the development of technological resources of potential value to English language teachers and learners. One such example is the word frequency highlighter, which allows you to submit a text of your choice for frequency analysis. Enter a text as in the screenshot below (www.lextutor.ca/vp/eng/) and when you click ‘submit’ the words in the text will be highlighted in different colours according to their frequency band.

If your interest is in identifying frequent academic words, an academic word highlighter can be found at http://www.nottingham.ac.uk/alzsh3/acvocab/awlhighlighter.htm. Enter your text and choose a level from 1 to 10. Level 10, as below, gives you all the frequent academic words.

The Academic Word List (AWL) highlighter also allows you to create a gap-fill by removing words of a specified frequency from the text you enter i.e. you choose a level between 1 and 10 for the gap-fill: level 10 will remove all the frequent academic words.

File produced at level 10

If you have had questions like these, this book is designed to help you to answer them by consulting corpora and corpus-informed literature. It is also designed to help you to generate and investigate similar questions. It is, however, important to keep corpora in perspective throughout this book. The argument presented here is that corpora are a resource and a reference source and, as in the case with all resources, pedagogic judgment is vitally important in determining how and when they are deployed to best effect.

The book does not assume prior knowledge or experience of corpus research; nor does it assume any technical expertise. Technology can relate contemporary corpus interfaces and corpus software are user-friendly and often include tutorial packages. The tasks in this book will help to familiarise teachers with publicly available user-friendly corpora such as the British National Corpus located at: http://corpus.lancs.ac.uk/

Figure 3.11 Academic word frequency highlighter

**Corpus Search**

1. Take an abstract from an academic article.
2. Predict five words which will be highlighted by the Academic Word Highlighter.
4. Check your intuition.
**Discussion**

(How) do you think you could use corpus tools such as the Academic Word Highlighter in your teaching?

How do you think your learners could make use of such tools autonomously?

**Collocation frequency**

Having stressed the importance of collocations and chunks at the outset of this chapter, it is important now to go beyond word frequency to consider collocation frequency and lexical chunk frequency. A challenge common to both these frequency calculations is identifying meaningful and useful units to include in the list. Computers can identify the most frequent 2-word, 3-word, 4-word combinations and so on with alacrity, but they cannot tell us whether these units have much inherent meaning – in other words they cannot distinguish between (applying our definitions) collocations and lexical chunks on the one hand and lexical bundles on the other; still less can they tell us which lexical items are the most important to learn. To illustrate the point about distinguishing between meaningful units and lexical bundles, we can return to our earlier observation (Shin and Nation 2008: 342) that the five most frequent 2-word lexical bundles in the Brown Corpus are: of the; in the; to the; on the; and the. These lexical bundles are hardly promising from a pedagogic point of view.

To address the challenge of focusing on maximally useful collocations (and excluding lexical bundles), Shin and Nation (2008) produced a collocation list from the spoken component of the British National Corpus based on six explicit criteria. Their collocation list is pedagogically motivated and intended for ‘beginning to lower intermediate students’. Shin and Nation’s (2008) six criteria are summarised below – we need to note first that they use the term ‘pivot word’ to describe the word which attracts the collocates (which we referred to earlier as a ‘node word’), e.g. *dire* is the pivot word of the collocations *dire straits*, *dire need* and *dire situation*.

Criteria for identification of collocations (Shin and Nation 2008):

1. Each pivot word was a word type rather than a lemma. That is, the different word forms ‘book’ and ‘books’ were treated as different pivot words.
2. The pivot word had to be a noun, a verb, an adjective, or an adverb. Adverbial particles like ‘up’ as in ‘get up’ were treated as pivot words because they were adverbs.
3. All the pivot words had to occur in the most frequent 1,000 content words (i.e. excluding from the list grammatical words such as prepositions and articles) of English according to the BNC spoken word frequency list by
Leech, Rayson, and Wilson (2001), available at http://www.comp.lancs.ac.uk/ucrel/bncfreq/flists.html. Thus, Shin and Nation (2008) decided to focus only on high frequency words. The working assumption was that the learning of the collocations should strengthen and enrich words students already know, not add an additional burden by adding unknown, lower frequency vocabulary.

Each collocation had to occur at least 30 times in ten million running words, i.e. the collocation should be frequent enough to get into the high frequency words of the language as if it were a single word.

Each collocation should not cross an immediate constituent boundary, i.e. it should be a complete grammatical unit such as a noun phrase, adjective phrase or preposition phrase. Thus ‘one of the’ would not be included, for example.

Different senses of collocations with the same form were counted separately. So, ‘looking up’ meaning ‘to improve’ was counted separately from ‘looking up’ meaning ‘to search for something’.

Shin and Nation (2008) note that criteria 3 and 4 above, in particular, demonstrate the pedagogic focus of their collocation frequency list, particularly the focus on the 1,000 most common words as pivot words. These criteria take us beyond identifying collocations based on an abstract definition of the term towards specifying which collocations should be included in the syllabus, a question on which we now focus. The top 20 most frequent collocations from the Shin and Nation (2008) list are listed below:

1. you know
2. I think (that)
3. a bit
4. (always, never) used to
5. as well
6. a lot of
7. (number) pounds
8. thank you
9. (number) years
10. in fact
11. very much
12. (number) pound
13. talking about {sth}
14. percent
15. I suppose (that)
16. at the moment
17. a little bit
18. looking at {sth}
19. this morning
20. (not) any more
Discussion

How useful do you think such a list is for teachers and materials writers?

While collocation frequency lists may be useful in informing the syllabus, in order to assess their value we need also to consider what challenges collocation present to learners. One aspect of the challenge identified by Nesselhauf (2005) is mother tongue interference which, in her study, accounted for about 50% of the learners’ collocation errors. While we might expect collocational competence to improve significantly as learners progress through the levels, there is evidence (Bahns and Eldaw 1993) that even advanced EFL students struggle with collocation and that, while some collocations are more ‘paraphrasable’ than others, a focus on collocations is still necessary even at advanced levels. A learner may, for example, be able to paraphrase ‘we made small talk’ by saying ‘we talked about unimportant everyday things’, but this will probably take longer to produce and also seem, to native speakers at least, less idiomatic and natural than the conventional collocation (whether the latter point should be of concern to us is a topic we deal with in Chapter 8).

Learners experience more difficulty, Ackermann and Chen (2013: 236) argue, in producing collocations than understanding them and need ‘a high level of collocational competence if [they] want to express themselves clearly and unambiguously’. They also refer to other research that establishes that learners find collocation challenging, summarising this research thus (Ackermann and Chen 2013: 236):

- Biskup (1992) showed that learners use inappropriate synonyms when producing collocations.
- Cobb (2003) found that even when learners use collocations correctly they over-rely on a small number of collocations.

Understanding the nature of collocation, and the problems learners have with it, is clearly a useful step towards teaching it, but understanding a linguistic phenomenon and being able to teach it are, as we know, not the same thing at all. Conzett (2000: 75) neatly sums up this potential gap between understanding collocation and teaching it: ‘After the initial Aha! feeling one has when realizing how much collocations come to bear upon language, the classroom teacher has to consider the question of just how to go about the explicit teaching of collocation’. Indeed, this comment by Conzett captures exactly how I felt after reading The Lexical Approach by Michael Lewis (1993): my first reaction was, ‘These insights into language are very exciting’. But my second reaction was, ‘OK, what do I do now as a teacher?’ An important part of the practical question was how to select which collocations to teach. Collocation frequency lists, as we have suggested in relation to words, are, for a number of reasons not enough:
• Collocation frequency lists don’t tell us which items are most difficult for learners or most useful for learners (for this we need learner corpora).
• Collocation frequency lists tell us little about how collocations might be grouped or sequenced for teaching purposes.
• Collocation frequency lists (for good reasons) cover only a fraction of the collocations in the language.

There are, however, specific suggestions in the literature as to how we might deal with collocations in pedagogically sound ways. Both Hill (2000) and Conzett (2000), for example, suggest that the strength of the collocation should be a selection criterion. The term ‘strength of collocation’ requires a little explanation. In these terms, the collocation good man would be regarded as a ‘weak collocation’ because good can collocate with a multitude of other words. Our earlier example of dulcet tones, however, would be regarded as a very strong collocation as dulcet does not collocate frequently with any other words. Between these two extremes we find a collocation such as severe winter, which would be regarded as a medium strength collocation because severe collocates with a reasonable, but limited number of other words, e.g. severe punishment, severe discipline, severe winter. Hill (2000) and Conzett (2000) suggest we focus on these medium strength collocations: though ‘medium strength’ is not a precise term, it suggests that we can focus on a limited number of collocates of the node word, i.e. the crucial point is that we aim to extend the learners’ collocational knowledge without overloading them. In practical classroom terms, this may involve building up a collocation table for the node word through elicitation, e.g.

<table>
<thead>
<tr>
<th>Node Word</th>
<th>Collocates</th>
</tr>
</thead>
<tbody>
<tr>
<td>severe</td>
<td>punishment</td>
</tr>
<tr>
<td></td>
<td>discipline</td>
</tr>
<tr>
<td></td>
<td>winter</td>
</tr>
</tbody>
</table>

The earlier decision to define a collocation as ‘two lexical words often found together’ is justified, I would argue, by the way it allows us to build such tables with the aim of facilitating learning.

Nation (2001) suggests the general criteria of the frequency and ‘range’ of the collocation (the tendency to appear in different genres and text types) as well as three other specific criteria:

1. We should pay most attention to frequent collocations of frequent words. Sometimes common collocations with common words are the ones that escape learners, e.g. do business; make a mistake; go to college; make a speech. I have, for example, lost count of the number of times learners have told me ‘I am
going to study a master’s course next year’ rather than ‘I’m going to do a master’s course’ (though you might argue that it doesn’t matter, a point to which we return in our later discussion of English as a Lingua Franca).

We should take into account the learning burden of a collocation to ensure that learners are presented with a manageable amount of information. The learning burden will be low if you can easily predict other collocates, e.g. from the collocation ‘raise prices’, the learners can probably predict other collocates of raise, e.g. ‘raise taxes’, ‘raise salaries’, ‘raise fees’.

Semantic unpredictability. It would not be obvious what ‘small talk’ means to someone who had never met the expression before, so we will probably want to draw learners’ attention to this collocation. ‘Baseball game’ and ‘football team’ are, however, semantically predictable and transparent so we can probably assume that most learners will find these unproblematic.

**Lexical chunk frequency**

As we have noted, it is not easy to generate automatically a list of meaningful lexical chunks, which presents a challenge for those wishing to compile a lexical chunk frequency list, particularly a pedagogically motivated list. However, Martinez and Schmitt (2012) took up this challenge by producing what they term a ‘phrase frequency list’ (based on the BNC) arguing that hitherto there had been ‘no principled way to prioritize the inclusion of such items in pedagogic materials, such as ESL/EFL textbooks or tests of vocabulary knowledge’. In relation to materials in particular they remark that ‘perusal of almost any EFL/ESL textbook or test yields a paucity of formulaic sequences targeted for explicit attention/noticing, and even for those that do occur, there does not seem to be much principled basis for selection’ (Martinez and Schmitt 2012: 301). Before we discuss their list further, we need to consider their definition of a ‘phrasal expression’ (Martinez and Schmitt 2012: 304):

A phrasal expression is hence defined as a fixed or semi-fixed sequence of two or more co-occurring but not necessarily contiguous words with a cohesive meaning or function that is not easily discernible by decoding the individual words alone.

Their list embraces then what we have termed collocations. Pedagogic concerns are clear in the three aims they outline for the phrase frequency list (Martinez and Schmitt 2012: 302):

- A guide for language learners and educators to include formulaic sequences in their learning and teaching, particularly for receptive purposes.
- A means of including formulaic sequences in tests that assess receptive L2 knowledge and receptive skills.
- An aid in monitoring vocabulary acquisition progress.
We noted in relation to the collocation frequency list that Shin and Nation limited their list to words in the top 1,000 word frequency band in order to facilitate learning; similarly, Martinez and Schmitt (2012) limit their list to the top 5,000 word frequency band. Martinez and Schmitt (2012) specify three main criteria applied in the composition of their phrase frequency list: frequency, meaningfulness and ‘relative non-compositionality’. The meaningfulness criterion allowed them to exclude frequent lexical bundles such as ‘as is the’ or ‘is of a’, while the relative non-compositionality criterion allowed them to focus on chunks whose meaning is not so easily deducible from the parts – we referred to this as semantic predictability in relation to collocations. As an example, Martinez and Schmitt (2012) note that the meaning of ‘at all times’ is more easily deducible from the meaning of the component parts than ‘at all costs’. By applying these three main criteria (and certain supplementary criteria), Martinez and Schmitt (2012) arrived at a list of 505 frequent phrasal expressions, noting that the vast majority of these were made up of the 2,000 most frequent words in the BNC. The most frequent phrases from this list are shown below (from: http://www.norbertschmitt.co.uk/resources.html):

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Frequency (per 100 million)</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. HAVE TO</td>
<td>83092</td>
<td>I exercise because I have to.</td>
</tr>
<tr>
<td>2. THERE IS/ARE</td>
<td>59833</td>
<td>There are some problems.</td>
</tr>
<tr>
<td>3. SUCH AS</td>
<td>30857</td>
<td>We have questions, such as how it happened.</td>
</tr>
<tr>
<td>4. GOING TO (FUTURE)</td>
<td>28259</td>
<td>I’m going to think about it.</td>
</tr>
<tr>
<td>5. OF COURSE</td>
<td>26966</td>
<td>He said he’d come of course.</td>
</tr>
<tr>
<td>6. A FEW</td>
<td>26451</td>
<td>After a few drinks, she started to dance.</td>
</tr>
<tr>
<td>7. AT LEAST</td>
<td>25034</td>
<td>Well, you could email me at least.</td>
</tr>
<tr>
<td>8. SUCH A(N)</td>
<td>23894</td>
<td>She had such a strange sense of humor.</td>
</tr>
<tr>
<td>9. I MEAN</td>
<td>23616</td>
<td>It’s fine, but, I mean, is it worth the price?</td>
</tr>
<tr>
<td>10. A LOT</td>
<td>22332</td>
<td>They go camping a lot in the summer.</td>
</tr>
<tr>
<td>11. RATHER THAN</td>
<td>21085</td>
<td>Children, rather than adults, tend to learn quickly.</td>
</tr>
<tr>
<td>12. SO THAT</td>
<td>20966</td>
<td>Park it so that the wheels are curbed.</td>
</tr>
<tr>
<td>13. A LITTLE</td>
<td>20296</td>
<td>I like to work out a little before dinner.</td>
</tr>
<tr>
<td>14. A BIT (OF)</td>
<td>19618</td>
<td>There was a bit of drama today at the office.</td>
</tr>
<tr>
<td>15. AS WELL AS</td>
<td>18041</td>
<td>She jogs as well as swims.</td>
</tr>
<tr>
<td>16. IN FACT</td>
<td>15983</td>
<td>The researchers tried several approaches, in fact.</td>
</tr>
<tr>
<td>17. (BE) LIKELY TO</td>
<td>15854</td>
<td>To be honest, I’m likely to forget.</td>
</tr>
<tr>
<td>18. GO ON</td>
<td>15610</td>
<td>He went on for a while before stopping for lunch.</td>
</tr>
<tr>
<td>19. IS TO</td>
<td>15232</td>
<td>Obama is to address the media this afternoon.</td>
</tr>
<tr>
<td>20. A NUMBER OF</td>
<td>15090</td>
<td>A number of concerns were raised.</td>
</tr>
</tbody>
</table>
Discussion

1. Are there any phrases you are surprised to see on the list? If so, try to account for their frequency.
2. How would you set about categorising such a list?
3. How do you think this list could be used in ELT?

As was the case with word and collocation frequency lists, domain-specific phrase frequency lists have begun to appear. Notable examples are the Academic Collocations List (Ackermann and Chen 2013) and the Academic Formulas list produced by Simpson-Vlach, R. and Ellis, N. (2010), both of which we discuss in Chapter 7.

While such pedagogically-oriented lexical chunk lists are a good starting point, I would argue that we still face formidable problems of selection with lexical chunks. As Moon (1997) notes, it is likely that proficient speakers of English have many, many thousands of such chunks at their disposal and there is no canonical list of chunks. Faced with a potentially bewildering choice, I believe that teachers and materials developers need, in addition to such lists, principled criteria to help inform their selections. This is particularly the case for teachers who have to make practical decisions on a daily basis, e.g.

- Are there any lexical chunks in this text which are worth highlighting for my students?
- Are there any lexical chunks I could teach to facilitate this speaking or writing task?

Evaluations of ELT materials, however, suggest that selection of lexical chunks for teaching is unsystematic – Martinez and Schmitt (2012) refer to the work of Koprowski (2005), Gouverneur (2008) and Hsu (2008) in this respect. Koprowski (2005: 322), for example, observed that in the three ELT coursebooks he researched the selection of ‘multi-word items’ seemed to be unsystematic:

Contemporary British coursebooks now routinely offer a generous and diverse mix of multi-word items … But while designers have been enthusiastic about adding chunks to the syllabus, the process of selecting items has been highly subjective and conducted without reference to corpus data.

The selection process for lexical chunks, Koprowski (2005: 328) concludes, appears to be guided by criteria other than the usefulness of the phrases themselves: ‘… it is evident that the designers did not structure their course around a body of useful lexical phrases, but rather, started in most cases with a theme, topic or structure and then considered items related to these basic concepts’. He deprecates
this practice on the grounds that it leads to the selection of too many lexical chunks of one type or lexical chunks grouped around particular structures. Lewis (2000) offered very broad criteria for the selection of chunks, highlighting the following types as particularly worthy of our attention:

- Discoursal responses, e.g. ‘that must have been awful’.
- Multi-word sentence adverbs, e.g. ‘to put it bluntly’; ‘personally speaking’.
- Discourse-organising sentence heads (see above), e.g. ‘The next point we consider is …’
- Prepositional phrases, e.g. ‘at the end of the day’; ‘in a way’.
- Lexical modality, e.g. the use of ‘tend to’ in examples such as ‘I tend to think’; phrases such as ‘there’s a chance that’, ‘there’s no doubt that’.

This represents a broad categorisation of chunks, however, rather than an itemised syllabus, and it mixes formal categories (e.g. prepositional phrases) with functional categories (e.g. discourse-organising sentence heads). At the same time, it could be argued that such selection criteria might be a useful point of departure for specifying a syllabus. Selection criteria are also suggested by O’Keeffe, McCarthy and Carter (2007: 71) who suggest a focus on pragmatic categories:

> We would argue … that it is in pragmatic categories rather than in syntactic or semantic ones that we are likely to find the reasons why many of the strings are so recurrent, and in the idea of chunks as frames that we will find the most pedagogically useful ‘handle’ on chunks for vocabulary teaching and learning. By ‘pragmatic categories’ we mean the different ways of creating speaker meaning in context. Such categories would include:

- discourse marking,
- the preservation of face and the expression of politeness,
- acts of hedging and
- purposive vagueness.

Thus far we have not discussed two kinds of chunk which tend to attract attention in the language classroom: idioms and proverbs. If we apply only our frequency selection criterion, they would probably not feature at all in view of Moon’s (2010: 201) observation that ‘… corpus studies of idioms and proverbs … suggest that most are infrequent, tending to occur mainly in journalism and fiction’. However, as we have noted, frequency can never be the sole determining criterion, and a focus on idioms or proverbs from time to time can have non-lexical benefits such as an insight into the target culture and cross-cultural comparisons of idioms and proverbs. We see here the need to set corpus insights in the broader pedagogic context.
Implications for the vocabulary syllabus

Having discussed words, collocations and lexical chunks separately, we need now to consider the notion of an integrated vocabulary syllabus. Given that some collocations and lexical chunks are more frequent than frequent individual words (O’Keeffe, McCarthy and Carter 2007; Shin and Nation 2008; Martinez and Schmitt 2012), it would seem that there is a strong case for moving towards an integrated vocabulary syllabus of words, collocations and lexical chunks, even from the earliest level:

... the vocabulary syllabus for the basic level is incomplete without due attention being paid to the most frequent chunks, since many of them are as frequent as or more frequent than single items which everyone would agree must be taught.

(O’Keeffe, McCarthy and Carter 2007: 46)

However, it is not just the a priori syllabus which should benefit from an integrated view of vocabulary. Corpus-informed teachers, arguably, will be better placed to make opportunistic decisions in the ELT classroom about what vocabulary to highlight from listening or reading texts or from classroom discussions.

Corpus Question

Look at the concordance of thing from the BNC Spoken Corpus and think about these questions:

1. How often (approximately) is the word thing used metaphorically as opposed to literally?
2. What chunks does thing appear in?
3. What pragmatic functions do these chunks have?

Figure 3.12 BNC concordance lines: thing
Conclusion

In this chapter we have discussed how corpus research has led to a different assessment of the relative importance of lexis and grammar, and a different view of the lexicon. This view accords greater importance to the most common words of English and to what we have chosen to term collocations and lexical chunks. We have also discussed the different aspects of work knowledge which have emerged from corpus research such as semantic prosody, semantic preference and semantic structure. Corpus frequency findings (in relation to words, collocations and lexical chunks) and, crucially, their potential implications for ELT have occupied a good deal of our discussion. It is apt to remind ourselves here of a point we made earlier: corpus findings should not dictate what is taught – they need to pass through a ‘pedagogic filter’, to use Timmis’ (2003) phrase. In other words, we are talking about a corpus-informed rather than a corpus-driven approach to teaching lexis.

References


Chapter 4

Corpus research and grammar

Introduction

In the context of our discussion of lexis in Chapter 3, we observed that corpus research has led to quite radical insights into the nature and role of lexis in communication and, most importantly for this chapter, fresh insights into the relationship between lexis and grammar. As we turn our attention to grammar, we will see that corpus research presents challenges to traditional orthodoxy in this domain too. Challenges to orthodoxy in the domain of grammar are, as Hughes (2010) observes, particularly sensitive since teachers often nurture deeply-held convictions about the importance of grammar, as well as strong attachments to traditional, prescriptive ‘rules’ about the ‘correct’ use of grammar. The descriptive and probabilistic ‘rules’ which often emerge from corpus research to cast doubt on the veracity of deterministic rules can, then, be unsettling for those attached to a straightforward correct/incorrect dichotomy in relation to grammar.

Drawing on Lawson (2001, cited in Barbieri and Eckhardt (2007)), O’Keeffe, McCarthy and Carter (2007) and Biber and Reppen (2002), we can summarise the main areas in which corpus research has contributed to our understanding of grammar. Corpus research has provided:

1 Information about the frequency of various grammatical structures both in general usage and across different registers.
2 Fuller, register-sensitive descriptions of how a range of grammatical features are used in writing and/or in speech.
3 A closer understanding of associations between grammatical features and the lexis which typically co-occurs with those features.

An alternative way to view the general tenor of corpus-based insights into grammar is offered by Timmis (2003), who suggests that corpus research into grammar offers both new perspectives on structures already described in traditional grammars and also new descriptions of grammatical features which have hitherto enjoyed little or no descriptive attention (often in the field of spoken grammar).
Corpus research has, for example, offered new (and challenging) perspectives on reported speech, an area that has always enjoyed attention in ELT materials (see below for a detailed discussion on this). In addition, corpus research, especially spoken corpus research, has often been able to show that structures previously thought peripheral, or even aberrant, are more frequent and systematic than previously supposed (McCarthy and Carter 1995). We look into spoken language research in more detail in Chapter 5, but we will take one illustrative example here, the ‘tail’ structure below, where we see a co-referential subject after the canonical clause structure. For example, in the example below, Bolton is co-referential with the subject it:

- It’s a funny population, Bolton is.

This structure has been covered only partially and spasmodically until relatively recently when spoken corpora have enabled fuller descriptions, showing how its use has to be related to affective and discourse concerns before it can be properly understood (e.g. Carter, Hughes and McCarthy 1998; Durham 2011; Timmis 2010).

In this chapter, we review some of the corpus-based descriptive findings in relation to grammar to illustrate how corpus research can shed new light on grammatical features. We focus particularly on three areas: the frequency of grammatical features, examples of insights from corpus-based descriptions of grammatical features, and the relationship between grammar and lexis. We also consider the potential pedagogical implications of the findings we have discussed in each of these areas.

**Grammar and frequency**

*Frequency findings from corpus research*

As we noted in relation to lexis, it seems uncontroversial to suggest that frequency of occurrence should be one criterion in the selection of grammatical items for the syllabus, and if we accept that frequency is an important criterion, we will be interested in corpus frequency counts in relation to specific grammatical structures. From a pedagogic point of view, we will be especially interested in apparent disparities between corpus-attested frequency of grammatical features and the prominence of these features in ELT materials (e.g. Mindt 1996; Shortall 2007). We consider below three examples of disparities between corpus frequency research and the prominence of grammatical features in ELT materials. We begin with two case studies of the presentation of grammar points in ELT materials by Biber and Reppen (2002) who focused on the treatment of pre-modification and of simple/progressive form in six ESL grammar books they surveyed for the purposes of their study. Our third case study looks at reported speech.
Corpus research and grammar

Grammar case study 1: Pre-modification with nouns and adjectives

Biber and Reppen (2002) deal first with the relative frequency of adjectival and noun pre-modification, arguing that the grammar books they surveyed focus on adjectival pre-modification to a disproportionate extent (‘relative frequency’ is an example of adjectival pre-modification; ‘grammar book’ is an example of noun pre-modification). They concede that in conversation nouns are pre-modified significantly more often by adjectives than by pre-modifying nouns. However, they point to a ‘dramatically different pattern of use’ (Biber and Reppen 2002: 202) in written registers, especially newspaper writing, where noun pre-modification is almost as frequent as adjectival pre-modification. In other words, there is a register difference in the type of pre-modification used, in this case a broad register difference between conversation and writing. Biber and Reppen (2002) go on to observe that, while noun pre-modification is not often taught, perhaps because it is regarded as unproblematic, it is more complex than it may seem at first sight. They point out that a variety of relationships are possible between the pre-modifying noun and the head noun, providing the following examples:

• glass windows, metal seat, tomato sauce (Noun 2 is made from Noun 1)
• pencil case, brandy bottle, patrol car (N2 is used for the purpose of N1)
• sex magazine:, sports diary (N2 is about N1)
• farmyard manure, computer printout (N2 comes from N1)
• summer rains, Paris conference (N1 gives the time or location of N2).

Given its frequency in written registers and its formal complexity, Biber and Reppen (2002) argue that noun pre-modification deserves more treatment in ELT materials than is currently the case. We can anticipate our discussion of learner corpora in Chapter 6 by remarking that frequency and usefulness are only two criteria: we also need to establish whether learners, or specific groups of learners, have any problems with this structure – it is certainly unlikely to cause problems receptively.

Discussion

How far do you agree that noun pre-modification deserves more attention in the classroom? What other factors do we need to take into account?

Grammar case study 2: Simple v. Progressive

The amount of attention typically given to progressive verb forms by grammar books is also challenged on frequency grounds by Biber and Reppen (2002), who
point out that simple verb forms are 20 times more common than progressive forms in conversation. We should note here that Römer (2005), in a survey of ELT textbooks, also argued that progressive verb forms were presented more prominently than their corpus-attested frequency would appear to warrant. This concentration on progressive forms can lead, Biber and Reppen (2002) note, to overuse of these forms by learners (certainly true in my experience). We can speculate that the reason for this focus on the progressive is that it is often perceived as a more difficult form than the simple aspect on the basis of anecdotal evidence, but it is nonetheless a striking example of an apparent disconnection between ELT practice and corpus grammatical frequency findings.

**Grammar case study 3: Reported speech**

As a further example of the disconnection between corpus-attested frequency and ELT materials, we can cite the prominence accorded to canonical forms of reported speech in materials. In this case, it is not the frequency of reported speech itself which is in question, but the frequency of specific devices used to report speech in conversation and in writing: a number of commentators (Lewis 1993; McCarthy 1998; Willis 2003) have pointed out that the ‘backshift rules’ so commonly taught in ELT materials, while they are often used in writing, rarely apply in conversation where a variety of other devices are used. Here again we see the suggestion that grammatical forms are register-sensitive (see the section ‘Corpus perspectives on reported speech’ for further discussion of reported speech).

We should note in passing the argument that certain grammatical forms which are common in conversation, e.g. the get passive, have received insufficient attention in ELT materials (McCarthy 1998). We return to the theme of the disconnection between corpus-attested frequency of structures in conversation and textbook treatment in Chapter 5.

**Corpus research into frequency: Pedagogic implications**

Corpus evidence, then, poses frequency-related questions about the grammatical items we include in the syllabus and the priority we accord them. In practice, Biber and Reppen (2002: 200) contend, ELT practitioners tend to rely on intuitive judgements rather than corpus-attested frequency of structures:

This disregard [of frequency] is often accompanied by the assumption that current pedagogical practice is more carefully grounded than a reliance on frequency. In practice, though, this is not the case. Rather, authors often make pedagogical decisions based on their beliefs about language use, in many cases without even acknowledging that decisions are being made …

It is as well to recall here that while corpus evidence poses questions about what grammar we teach and how we describe it, the corpus-informed approach we
discussed in the introduction means that the evidence should not dictate what we do. Alongside the frequency of a structure, we need to take into account its difficulty and usefulness for a specific group of learners. We also need to acknowledge Widdowson’s observation that ‘frequency does not ... reflect prototypicality, the relative cognitive salience of words and structures’ (Widdowson 2003: 87). To take a lexical example, it may well be that the metaphorical use of bet is more common than its literal meaning, but we may only understand the metaphorical meaning through the literal meaning: it is the literal meaning which is cognitively salient. In the case of grammar, the present simple may be massively more common than the present progressive, but a teacher may decide to devote more attention to the progressive because it presents more problems for a specific group of learners and/or because it is easier to illustrate through classroom actions. Alternatively, noun pre-modification may be frequent, but is it likely to cause any problems for learners? There is not necessarily a direct match between ‘real-life’ communicative importance and classroom importance: Widdowson (1991: 21) suggests that certain (words and) structures may be seen as ‘pedagogically core or nuclear’ because they are ‘catalysts which activate the learning process’ irrespective of their frequency.

These are important caveats in relation to frequency as a selection criterion, but, as Gavioli and Aston (2001) argue, corpus evidence can play an important role in prompting us to reflect on and justify our choices as to the sequence in which we teach grammatical structures. Specifically, we need to ask ourselves questions such as:

• Are we teaching a structure because it is frequent and/or demonstrably useful for our learners or simply because it is featured in the course materials (and always has been)? Are we blindly following convention? Biber and Conrad (2010) have argued that teachers often assume there is a rationale for the grammar syllabus based on ‘secret’ knowledge held by course designers about grammatical frequency and difficulty. In reality, the syllabus seems to be based on tradition.

• Are we teaching the structure because it is frequent and/or demonstrably useful for our learners or simply because it lends itself to systematic presentation and practice? We should bear in mind here Thornbury’s (2000) memorable phrase ‘Grammar McNuggets’ to describe structures which are easily packaged in the classroom without necessarily being ‘nutritious’. In my role as a tutor on pre-service teacher training courses, for example, I observed countless lessons on the structure used to when used to discuss past habits. Was it selected because it is a particularly frequent and useful structure or because it lends itself to PPP (present–practise–produce) presentation with engaging personalised practice?
Discussion

1. Examine the grammar syllabuses/contents pages of two or three intermediate coursebooks. How similar or different are they? Can you see an overall rationale for the syllabus?

2. What criteria would you apply in selecting items for a grammar syllabus?

Corpus perspectives on grammatical description

Corpus evidence does not, however, only raise questions about the frequency of grammatical items in relation to their prominence in the syllabus; it also questions the completeness and/or accuracy of existing pedagogic descriptions of how grammatical features are used. A number of commentators have referred to the differences between grammatical descriptions (or prescriptions) presented in coursebooks and grammatical descriptions based on corpus evidence. Mindt (1996: 232), for example, speaks of ‘a kind of school English which does not seem to exist outside the foreign language classroom’. Similarly, Römer (2006: 125–126) speaks of ‘considerable mismatches between naturally-occurring English and the English that is put forward as a model in pedagogical descriptions’. Römer (2006) also cites specific research which shows that corpus-attested use of modal verbs, if-clauses and the present perfect differs significantly from the use of these structures as presented in ELT materials. To illustrate more specifically how pedagogic descriptions (and prescriptions) of grammatical features can be said, on the basis of corpus evidence, to offer only partial and/or unrepresentative accounts of particular items, let us take three illustrative examples from the literature: reported speech, conditionals, and the use of though as a conjunction and as an adverbial.

Corpus perspectives on reported speech

We have already noted that the pedagogic treatment of reported speech has come in for criticism by a number of commentators: indeed, it has become something of a punch-bag for corpus heavyweights, (and sometimes even a punch-bag for corpus lightweights keen to flex their critical muscles). However, our initial observation was that a particular form of reported speech, ‘backshift’, was over-represented in ELT materials in relation to frequency. Here we focus on the adequacy of pedagogic descriptions of reported speech, drawing mainly on Barbieri and Eckhardt (2007), who compared reported speech as presented in ELT materials with corpus-attested reported speech. Their analysis is based on two corpus studies: 1) Eckhardt (2001) which used a 3.3 million word corpus of newspaper writing and a 2.1 million word corpus of everyday conversation drawn from the American English component of the Longman Spoken and Written English (LSWE); and 2) Barbieri (2005) which was based on extracts from the
American Conversation component of the LSWE corpus and the three components of the TOEFL 2000 Spoken and Written Academic Language corpus: service encounters, study groups, and office hours.

ELT materials, as Barbieri and Eckhardt (2007) observe, tend to focus on the ‘backshift’ rules of reported speech involving transformations of sentences in direct speech to indirect speech. This focus on backshift rules, rules which are rarely applied in speech (McCarthy 1998), they argue, leads to a neglect of other common structures which are commonly used to report speech in spoken language. They point in particular to the neglect of what they term the ‘new quotative verbs’, e.g. be like, go, and be all, noting that these verbs are used across different spoken registers (conversation; service encounters; study groups; office talk). These three quotatives are exemplified below with excerpts from COCA:

- And **she was like**, “I, I can dance” **I was like**, “Girl, if you don’t shut up, forget dancing. Just stand right there and do what you do”.
- “You wouldn’t know I went to Yale talking like that, I suppose,” he said, and **I was all**, “Yale?” and he said, “Yeah, I have a law degree that I don’t use anymore”.
- he **goes**, “No, it’s pretty – it’s pretty lonely. So – pretty darn lonely the whole way”.

Barbieri and Eckhardt (2007) observe that a further quotative device, the use of the past continuous for the reporting verb is also neglected (see McCarthy 1998), e.g. (extract from COCA):

**I was just saying** to Christine the other day how I saw a picture of her recently and I just thought, she’s turning into such a beautiful woman. I never – I’ve always thought she …

Barbieri and Eckhardt (2007) conclude that ‘far from being the “monolithic” linguistic phenomenon portrayed by ESL/EFL grammar textbooks, RS [reported speech] displays significant variation across different registers’.

As our focus is on corpus linguistics with specific relation to ELT, it is particularly interesting for us that Barbieri and Eckhardt (2007) provide concrete suggestions as to how reported speech should be taught in the light of these corpus findings. I have summarised their main suggestions below:

1. Reported speech should **not** be taught through exercises where direct speech is transformed to reported speech (Stranks [2013] also criticises the transformation approach to reported speech on the grounds that when we produce reported speech we do not naturally go through a transformation process: we go directly to reported speech).
2. Indirect reported speech should be taught in the context of newspaper writing, where it is quite common, whereas direct speech reporting should be taught in the context of conversational spoken registers.
The reporting verbs *say* and *tell* should be prioritised for indirect reported speech and attention drawn to other frequently used reporting verbs, e.g. *announce, report, agree, claim, warn.*

For DRS (direct reported speech), we should teach the standard, unmarked quotative ‘say’, but provide positive evidence of at least two widely used, non-traditional quotatives: *be like* and *go.*

We should raise awareness of the situational, discourse-pragmatic and socio-linguistic factors which condition the quotative use of *be like, go and be all.*

We can see in the specific pedagogic recommendations above two recurring implications of corpus research for ELT pedagogy: the need for a register-sensitive approach to grammar teaching and the potential compatibility of awareness-raising approaches with corpus-based grammatical descriptions.

Discussion

How far do you agree with the teaching recommendations for teaching reported speech by Barbieri and Eckhardt (2007)? Can we distinguish between teaching grammatical items for production and reception?

Conditionals, corpora and pedagogy

The next god of the grammatical pantheon to be slain by corpus linguists is the pedagogic treatment of conditionals, which has come under fire on the grounds that the rules traditionally presented in ELT materials are a considerable oversimplification, if not misrepresentation of reality. For our case study of conditionals, we draw here on two pedagogically-oriented corpus studies: Frazier (2003) and Jones and Waller (2011). For his study, Frazier (2003) drew on three corpora: the Brown Corpus, the Santa Barbara Corpus of Spoken American English and the Michigan Corpus of Academic Spoken English (MICASE). For their study, Jones and Waller (2011) extracted 250 concordance lines from the British National Corpus.

One finding of immediate relevance to us is that, while conditionals are often referred to as ‘if-clauses’ in ELT materials, there is evidence to show that this term is somewhat misleading (Frazier 2003: 443):

Contrary to the way these (conditional) structures are represented in ESL/EFL textbooks – as would-clauses adjacent to conditional clauses with if – these corpus data indicate that would-clauses in counterfactual/hypothetical environments occur more often quite distant from or entirely without any corresponding if-clauses.
Similarly, Willis (1990, cited in Jones and Waller 2011) found that the majority of hypothetical uses of *would* in the Cobuild Corpus had no associated if-clause. Indeed, in relation to ELT materials, Frazier (2003: 446) refers to ‘… a marked neglect of discussion on hypothetical or counterfactual would-clauses that do not occur in the same sentence as their supposed conditional if-clauses …’.

In conversation, an if-clause can be uttered by one speaker and the result-clause by another or, as in the example below, turns may intervene before the speaker adds the result clause to the condition. In the extract below, for example (Frazier 2003: 443–446), the first example of a conditional is canonical, but the second is only completed over a span of several turns:

Brad: One thing that comes to mind is, is the, well the question of meeting quarterly, is that enough. Meaning having, you know, the-

Phil: **It would be enough if committees were to meet.**

Brad: **If committees yeah.** And that’s the-

Phil: But-

Brad: That—that’s—that was Paul’s, uh, desire.

Phil: And I think.

Brad: Was uh-

Phil: that’s what we wanna talk ab- we’ll talk about that there too. And say look-

Brad: Mhm.

Phil: You know, is quarterly—you know, **we- would we need to meet every other month.**

Brad: Mhm.

It has long been customary in ELT materials to identify four main types of conditional, as exemplified below:

Zero conditional: If you pay peanuts, you get monkeys. [If + present simple + present simple]

First conditional: If you read this book, you’ll know more about corpus linguistics. [If + present simple + will + bare infinitive]

Second conditional: If I had time, I’d write a novel. [If + past simple + would + bare infinitive]

Third (past) conditional: If I’d known you were coming, I’d have baked a cake. [If + past perfect + would + have + past participle]

While these four types are long-established in ELT materials, Jones and Waller (2011) draw attention to the usefulness of Maule’s (1988: 122) classification of conditionals with its focus on the distinction between real and unreal conditionals, as illustrated below, particularly as it allows for real past conditionals (example b in the table):
Table 4.1 Conditionals typology (adapted from Maule 1988)

<table>
<thead>
<tr>
<th>Type of conditional</th>
<th>Example</th>
<th>Forms used</th>
</tr>
</thead>
<tbody>
<tr>
<td>a Real non-past</td>
<td>If you’re having another, get me one</td>
<td>If + present continuous, present simple</td>
</tr>
<tr>
<td>b Real past</td>
<td>If you didn’t want to do it, why did you agree to do it?</td>
<td>If + past simple, past simple</td>
</tr>
<tr>
<td>c Unreal non-past</td>
<td>If I knew, I would tell you.</td>
<td>If + past simple, would + infinitive</td>
</tr>
<tr>
<td>d Unreal past</td>
<td>If we had lost the match, I would have resigned</td>
<td>If + past perfect, would + have + past participle</td>
</tr>
</tbody>
</table>

Jones and Waller (2011) note that ‘Carter and McCarthy (2006: 749-50) found many samples of real non-past and real past conditionals in the Cambridge International Corpus, such as the following:

- If + going to, going to, e.g. ‘If you’re going to buy a house, then you’re going to need a lot of money’ (real non-past).
- Past simple, if + past continuous, e.g. ‘They always took the dog with them if they were going anywhere’ (real past).

Referring to their own data from the BNC, Jones and Waller (2011) also note, among many others they cite, the following possibilities which are not covered by canonical descriptions:

- If + past simple, past simple: ‘… if you wanted to know the answer … you had to keep zapping from channel to channel’.
- If + present simple, ‘would’: ‘If two members of staff happen to fall in love and decide to marry it would be churlish to be appointing blame’.

While it is true that at advanced levels, learners are occasionally let into the guilty secret that there are ‘mixed conditionals’ which do not conform to the canonical rules, in general, corpus studies point to two main perceived inadequacies of the treatment of conditionals in ELT materials:

1. The co-occurrence of ‘if’ with conditional forms is exaggerated.
2. The combinations of verb forms found in condition and result clauses are considerably over-simplified.

We now need to consider practical teaching proposals for the conditionals which take into account this corpus evidence. A number of concrete suggestions are made by Jones and Waller (2011) who begin, however, by acknowledging that there are reasoned defences of the status quo in terms of presenting the conditionals in ELT materials, acknowledging, for example, Ur’s (1989) argument that, while the first,
second and third conditionals may not be the most frequent conditional forms, they
deserve particular attention because they are difficult for learners as their meaning
is non-transparent. Jones and Waller (2011) also acknowledge Soars and Soars’
(1996) argument that the four canonical conditionals can be seen as the root
conditionals from which the various other ‘mixed conditionals’ are derived. The
objection to this argument that is raised by Jones and Waller (2011: 29), however,
is that, while it might be sensible to start with the four canonical forms, ELT
materials rarely go beyond what might seem to be sensible starting points:

These arguments would be fine if the four traditional patterns were seen as a
starting point for examining real and unreal uses of conditionals. Unfortunately,
the reality is that they are not normally presented in this way and instead an
oversimplification is all most learners are exposed to. This does not seem to
be a fruitful way to help learners cope with the variety of conditionals in
actual use.

Jones and Waller (2011) go on to make a number of suggestions as to how the
conditionals could be taught. Their main points are summarised below:

1. We should spend more time on real conditionals (and their different patterns)
   than on unreal conditions on the grounds of frequency.
2. We should raise learners’ awareness of conditionals which do not contain an
   if-clause.
3. Trainee teachers need to develop an awareness of conditionals beyond ‘The
   big four’, perhaps through discussion of corpus data.
4. One way for learners to study conditionals might be through reviewing
   examples in a pedagogic corpus, i.e. a corpus built up from the learning
   materials used in class (see Chapter 1).

We see in these proposals by Jones and Waller (2011), as we noted in relation to
Barbieri and Eckhardt’s proposals for teaching reported speech, that corpus
research is accompanied by a pedagogic emphasis on awareness-raising. While
we have been discussing here how conditionals might be presented to learners, we
should keep in mind that this will not happen unless teachers and materials writers
see beyond the four conditionals: it is for this reason, Jones and Waller (2011)
argue, that awareness-raising tasks focusing on the conditionals should be used on
teacher training courses.

**Discussion**

Do you agree with the teaching proposals for conditionals suggested by Jones
and Waller (2011)? Why?
Not such a simple word though

Our third and final case study of differences between corpus findings and treatment in ELT grammar books is briefer than the previous two but, I hope, instructive nonetheless: it concerns the use of though. Note in the screenshot below how the search can be restricted to a specific component of the corpus, in this case the spoken component.

What do you notice about the spoken use of though in the examples below from the BNC?

While the discourse marker though is a well-established item in ELT materials, it is often presented as a conjunction used in either initial or medial position, as in the examples below (from the BNC):

1 **Though** he devoted six years of his life to the RPF, de Gaulle virtually ignored it in his memoirs.
2 They didn’t say where it would be held though it’s thought the Soviet leader will visit Washington from the thirtieth May until June third.

Spoken language research, however, presents a strikingly different picture of the use of though in conversation. Conrad (1999; 2004) makes three interesting points about the use of though in conversation:

1 It is often used clause-finally as a linking adverbial, e.g.
   S1: Oh is he coming?
   S2: Think so yeah but he was ill though yesterday.
2 Though is one of the most common linking adverbials of contrast/concession in conversation.
3 Though can be used interactively between speakers, e.g.
   S1: It’s not nice.
   S2: It’s funny though.
   (Example from the BNC)
4 Though can play an important interpersonal role in softening disagreement, as we can see from the example above where S2 does not challenge S1’s evaluation of a joke, but rather appears to draw attention to a different aspect of it. This use is described by Conrad (2004: 72) thus:

   …an important function of though is apparent when it is used during exchanges between speakers. Though provides a means of disagreeing in a less direct way than with but or however. The linking adverbial seems concessive in meaning, as though the second speaker is not contradicting the first, but just adding information that needs to be taken into account. The disagreement is softened …
Figure 4.1  BNC (spoken) concordance lines: though
Given the frequency and apparent value of *though* as a linking adverbial in conversation, the results of Conrad’s (2004: 73) survey of four ESL textbooks are perhaps surprising: ‘Only one of the four [ESL textbooks] covers the use of *though* as linking adverbial at all, and that book lists it only as showing contrast, not concession. None of these books have an example of *though* to soften disagreement’. It is worth noting, however, that the interactive use of *though* as a linking adverbial *is* covered in *Touchstone*, written by McCarthy, McCarten and Sandiford in 2005, whose materials are informed by the North American component of ICE. It is also particularly interesting for our purposes that the authors of *Touchstone* have opted for an explicit PPP approach with *though*, involving the overt explanation and controlled practice often absent from awareness-raising approaches. While we saw in relation to the studies by Barbieri and Eckhardt (2007) and Jones and Waller (2011) a suggestion that a different (i.e. corpus-informed) approach to language description might imply a different methodological approach (awareness-raising), McCarthy (2004) outlines a different stance, arguing that precisely because corpus-based grammatical description can produce insights which challenge orthodoxy, methodology should remain reassuringly familiar. We should note here Willis’ (2003) argument that the approach we adopt should be related to the nature of the point we are teaching: if it is essentially a matter of form, then explicit presentation is likely to be suitable; if more subtle conceptual distinctions are involved, then a more gradual awareness-raising approach is likely to be suitable. Here we could argue that the position of *though* is a matter of form and, therefore, amenable to a PPP approach.

Below are further examples of studies which show clear differences between grammar in ELT materials and corpus-attested use of grammar:

- Shortall (2007) observed that the active forms of the present perfect are over-represented at the expense of passive and modal perfect forms. He also noted that the time adverbials *since* and *for* are over-represented compared with other time adverbials such as *now*.
- Mindt (1996), in a review of ELT coursebooks used in German schools, pointed out that *going to* for the future is taught before *will* even though *will* is a far more frequent form and easier to learn. He also draws attention to the omission of *shall*, despite the fact that it is common in first person forms.

Such studies, Ruehlemann (2008: 16) claims, ‘reveal a clear mismatch between the corpus evidence and what is covered in the textbooks: either the textbooks represent less frequent features at earlier stages than more frequent ones, that is, in an order which disregards their importance in actual discourse, or they fail to represent them at all’.
Grammar and lexis: Connections

In this section we continue to refer to an illustrative sample of corpus-based descriptive insights, but we are concerned with the ‘bigger picture’: what does the cumulative effect of these insights say about the relationship between grammar and lexis? There have been strong claims that corpus evidence points to a closer relationship between grammar and lexis than has traditionally been supposed. Both Römer (2009) and Biber and Conrad (2010) refer to the weight of evidence supporting such claims:

- Corpus studies, based on large collections of authentic text from a range of different sources, have provided massive evidence for the interdependence of lexis and grammar (or vocabulary and syntax): Römer (2009: 141).
- Corpus-based research has consistently found that there are actually strong associations between grammatical structures and the words used with them: Biber and Conrad (2010: 4).

There are two ways to examine the relationship between lexis and grammar: we can begin with lexis and look at the structures a particular word tends to co-occur with; or we can start with a structure and look at the kind of words it typically attracts.

The lexis and grammar connection

We begin by looking at the lexis–structure relationship: Biber and Conrad (2010: 4), for example, note that certain verbs, when they are used, tend to be used in the progressive aspect. The examples they cite are: bleeding, chasing, shopping, starving, joking, kidding, and moaning (to clarify the distinction we made above, we need to keep in mind that these are not the most common verbs used in the progressive aspect). To take a further example of the lexis to grammar relationship, Biber et al. (1999: 459) list verbs which are used 80% in the present tense and verbs which occur 80% in the past tense:

- Verbs occurring over 80% of the time in the present tense – bet, doubt, know, matter, mean, mind, reckon, suppose, thank.
- Verbs occurring over 80% of the time in the past tense – exclaim, eye, glance, grin, nod, pause, remark, reply, shrug, sigh, smile, whisper.

As Biber et al. (1999: 459) observe: ‘The present tense is strongly associated with denoting mental and logical states … while the past tense is strongly associated with verbs denoting events or activities …’. This comment brings to mind Stubbs’ (1995) remark that many corpus insights are obvious when you know them.

Common lexis–structure associations are also noted by Hunston and Francis (1998). They point, for example, to an association between verbs expressing
like/dislike and cleft constructions, e.g. ‘What I loved about Gloria was her talent and her independence’. You can check this by carrying out the following search at: http://corpus.byu.edu/bnc/

Note how the part of speech is specified by using the PoS button to produce the string:

What I [vv*] about

In a more extended example, Hunston and Francis (1998) refer to the structure ‘V by -ing’, e.g. ‘I’ll start by giving an example’. They note that the 20 verbs which are commonly used in the pattern, fall into two meaning groups:

Most of the verbs with this pattern fall into two meaning groups. They mean either ‘start’ or ‘finish’ (begin, close, end, finish, finish off, finish up, open, start, start off, start out); or ‘respond to or compensate for something’ (atone, compensate, counter, react, reciprocate, reply, respond, retaliate). Two other verbs (live and profit) are to do with gaining resources (Hunston and Francis 1998: 211).

It is also possible to look at the relationship between a particular sense of a word and the grammatical patterns with which it typically occurs. We begin with Hunston, Francis and Manning’s (1997: 210) observation about the word face: they note that in its more metaphorical sense of aspect, it tends to occur in the pattern: the + adj + face + of, citing the following examples:

• the acceptable face of the Cambodian government
• the ugly face of Western authoritarianism.

A further example of this sense/pattern relationship is provided by Hunston and Francis (2000: 255–256) who note that the verb reflect has different complementational patterns according to the sense in which it is used:

• V + N (to reflect light)
• be V-ed (sth. is reflected in a mirror)
• V + Prep (she reflected on the relationship between pattern and meaning).

Finally in this section, with the aid of an example from O’Keeffe, McCarthy and Carter (2007), we can see a connection between sense/pattern relationships and semantic prosody (see Chapter 2). O’Keeffe, McCarthy and Carter (2007: 103) observe that the preposition on occurs frequently with the words bordering and bordered and point out that bordering on and bordered on tend to be figurative rather than literal uses of border; just as importantly, the semantic prosody of the figurative use is negative, e.g. bordered on the ridiculous. As we are interested in potential pedagogic applications of research, we should take note of the argument
Figure 4.2 BNC search: What I [verb] about
that, ‘The grammatical patterns entail semantic patterns that learners of the language also need to know’ (O’Keeffe, McCarthy and Carter 2007: 104).

From the collocation list below (for bordered on) note how inferring semantic prosody involves qualitative work. No single negative word is especially frequent, but there are a number of words with a negative sense:

![BNC collocation search: bordered on](image)

**Discussion**

How far do you think lexis and grammar connections as exemplified above can be exploited for teaching purposes?

**The grammar and lexis connection**

To illustrate the reverse relationship – grammar/lexis – we will look at four examples of the association between grammatical structures and lexis:

- present perfect
- the verb + infinitive structure
- present tense
- past tense.

We begin with the most common verbs which occur in present perfect aspect (Biber *et al.* 1999: 464): been, had, got, gone, done, made, seen, come, said, taken.

A further example of the grammar of the grammar/lexis link is provided by Biber and Conrad (2010: 4) who observe that verbs which occur in the v + inf structure can be divided into four sense groups:
Corpus research and grammar 75

- want/need verbs: hope, like, need, want, want NP, wish
- effort verbs: attempt, fail, manage, try
- begin/continue verbs: begin, continue, start
- ‘seem’ verbs: appear, seem, tend.

Conrad (2010: 230) points to a semantic link between verbs commonly used in the present tense:

The verbs most strongly associated with present tense convey mental, emotional and logical states. Many of these are used in short, common expressions in conversations expressing the speaker’s mental or emotional state:

- It doesn’t matter
- I suppose.

Similarly, Conrad (2010: 230) notes a semantic link between verbs commonly used in the past tense: ‘The verbs most strongly associated with past tense … convey events or activities, especially body movements and speech’.

A further aspect to take into account in relation to grammar–lexis links is that such links can be genre-sensitive, e.g. Conrad (2010: 229–230) observes that ‘verb + that-clause structures in academic prose are used to report previous research, often with non-human entities acting as the subject (e.g. …reports suggest…).

The grammar/lexis and lexis/grammar connection:
Implications for pedagogy

We have already referred to Biber and Reppen’s (2002) observation that little thought seemed to have been given to the lexis used with the structures they examined in their survey of ELT materials. Indeed, Stranks (2013) criticises ELT materials writers for ‘random lexicalisation of structures’, i.e. making insufficient attempt to exemplify grammatical structures through the lexis which typically co-occurs with these structures. A slightly different perspective on the lexicalisation of structures is taken by Biber and Reppen (2002) who focus on whether verbs which are highly frequent in general terms (rather than in relation to the specific structure) are used to exemplify two basic structures: present simple and present progressive. They identify the 12 most common verbs in conversation, noting that these verbs account for 45% of the use of lexical verbs in conversation. Their survey of six low intermediate coursebooks revealed, however, that seven of these 12 common lexical verbs were not used at all to exemplify present simple and present progressive. This seems like a missed opportunity to reinforce key lexis while practising core grammatical structures.

We need now to consider the potential pedagogic benefits of raising learners’ awareness of what Hunston, Francis and Manning (1997) term ‘pattern
grammar’, i.e. the link between the meaning of words and the structures they typically co-occur with. We also need to consider how awareness of such patterns might best be raised. Fortunately, we can turn to Hunston, Francis and Manning (1997), enthusiastic proponents of a focus on patterns in the classroom, for advocacy of the benefits of such a focus and concrete suggestions as to how it might be achieved. Hunston, Francis and Manning (1997) outline four potential benefits of a pattern grammar focus: promoting understanding; promoting accuracy; promoting fluency and promoting flexibility. An awareness of patterns, they argue, can promote understanding by helping learners to guess unknown words from context, exemplifying this claim thus: a learner may not know the word *hail* in the sentence below:

This work has been **hailed** as an important step in trying to understand how life evolved.

However, if the learner is familiar with the pattern ‘V n as n’ (passive ‘be V-ed as n’) as used with verbs such as *announce, classify, describe, interpret, label, portray, proclaim, regard*, the process of guessing from context will be facilitated. Accuracy is promoted by learning pattern grammar, it is argued, as learned patterns potentially lead to error-free strings of text. It is this ability to produce strings of text automatically that also underpins the claim that learning patterns improve fluency (we made similar claims in relation to lexical chunks in Chapter 2). Finally, learning patterns can lead to flexibility in two ways: the acquisition of synonyms which express the same idea in the same pattern, and the acquisition of different patterns to express the same idea. Hunston, Francis and Manning (1997) exemplify the benefit of the latter thus, showing how five different patterns can be used to express essentially the same meaning.

- She liked the idea.
- She warmed to the idea.
- She thought the idea a cracker.
- She considered the idea brilliant.
- She regarded the idea as brilliant.

In terms of practical applications in the classroom, Hunston, Francis and Manning (1997) suggest three ways in which awareness of patterns can be raised:

- Asking learners to identify given patterns in texts used in reading classes.
- Giving (short) lists of words that share a pattern and asking learners to identify for themselves the meaning groups.
- By asking them to look and listen for a particular pattern over a given period of time, noting the words it is used with (if learners have access to English outside the classroom).
We could add here that rather than simply *giving* short lists of words that share a pattern, we can *elicit* words from our learners which share the same pattern. Elicitation might temper the potential danger of lexical overload as it draws on the existing knowledge of (some learners). One way of doing this kind of elicitation is to ask learners to replace selected words in a text without changing either the meaning or the grammar of the text as in the following example.

Think of alternatives for the underlined word in the sentence which keep approximately the same meaning without changing the grammar of the sentence:

The university has been *turned* into a sausage factory.

**Conclusion**

In conclusion, we have seen that corpus research has contributed to our understanding of grammar in a number of important ways. We now have more accurate information on the frequency of grammatical structure and more understanding of how this frequency varies according to the register we are considering. This kind of information provides, at the very least, food for thought when we consider the content and sequence of a grammar syllabus. We have seen that corpus research has led to fuller descriptions of certain grammatical features which sometimes challenge traditional pedagogical descriptions. It seems reasonable to suggest that these corpus-based descriptions should at least inform the grammatical descriptions we give in the classroom, though we acknowledge that there is sometimes a need to simplify and perhaps over-generalise to arrive at ‘workable’ pedagogic rules. Finally, the links between grammar and lexis, which corpus research has highlighted, offer promising avenues for more integrated teaching of grammar and lexis. The potential of corpus research to contribute to grammar teaching is neatly summarised by Conrad (2000: 549):

In the 21st century, I will argue, three changes prompted by corpus-based studies of grammar have the potential to revolutionize the teaching of grammar:

1. Monolithic descriptions of English grammar will be replaced by register-specific descriptions.
2. The teaching of grammar will become more integrated with the teaching of vocabulary.
3. Emphasis will shift from structural accuracy to the appropriate conditions of use for alternative grammatical constructions.

What is particularly interesting about this quotation, I would argue, is that Conrad initially talks about the ‘potential’ of ‘corpus-based’ studies, but then goes on to make three firm predictions: at the time of writing, I believe that we are still very much in the realm of potential rather than fulfilment of these predictions.
Corpus Search

1 Consult the BNC to find examples of the present perfect passive. See Figure 4.4 for a search tip:

![BNC search interface](image)

*Figure 4.4 BNC search: has been*

2 Can you find any semantic links between the verbs used in this form? If so, do you think the links could be exploited pedagogically?

---

**Corpus Search**

1 Consult the BNC to find nouns used as subject with the present perfect passive. Search tip: None! Work it out yourself this time!

2 Can you find any semantic links between the nouns which co-occur with the present perfect passive?
References


Jones, C. and Waller, D. (2011) If only it were true: the problem with the four conditionals. *ELT Journal* 65/1: 24–32.


Chapter 5

Spoken corpus research

Introduction

We have already referred a number of times to the findings of spoken corpora, particularly in the general chapters on corpus findings in lexis and grammar. In this chapter, spoken corpus research becomes our main focus for three main reasons:

1. Spoken corpus research is a relatively recent phenomenon and, while an interest in spoken language is not in itself new, it is the advent of discreet modern recording technology that has facilitated insights into spoken language hitherto difficult to access. We need, then, to be able to assess the relevance of these spoken corpus findings for contemporary ELT, particularly in an era when CLT and related methodologies such as TBL, which prioritise speaking skills, dominate the ELT landscape (McCarthy and Carter 2001).

2. As we noted in relation to corpus grammatical findings in Chapter 4, such findings can challenge long-held convictions about ‘correct’ language use and the nature of grammar. This seems to be particularly the case with spoken grammar, which can often be dismissed as slang, dialect or non-standard despite the evidence from spoken corpora that many spoken language forms so regarded are in fact systematic and used across the social spectrum.

3. As we shall look at in some detail in the next chapter, discussion of the relevance of spoken corpus findings to ELT has considerable significance for the contemporary debate about the role of the native speaker in a world where English is increasingly used in international contexts (Timmis 2012).

The chapter begins with a brief survey of spoken corpora available (but see also chapters 6 and 7 for learner spoken corpora and ESP spoken corpora respectively). The chapter then presents a review of spoken corpus findings in the areas of lexis, grammar and discourse which, it is argued, are of potential relevance to ELT. Such a review cannot be comprehensive: the aim is to consider a range broad enough to indicate the general tenor of such findings and to highlight issues in relation to the desirability and feasibility of teaching (native speaker) spoken
language. A further caveat is in order: the division of the review of findings into lexis, grammar and discourse is an organisational convenience, but we need to keep in mind that corpus research has tended to show that, in both spoken and written language, the boundaries between these three areas are permeable. Finally, after the review of research findings, we discuss the implications of spoken corpus research for ELT materials and methodology.

**Examples of spoken corpora**

While it is perfectly possible, as we discussed in the introduction, to build your own spoken corpus (see also O’Keeffe, McCarthy and Carter 2007: Chapter 1), it is useful to be aware of the kinds of spoken corpora that have already been constructed, especially if these are also freely available. Here we consider three types of spoken corpus: 1) spoken components of large general corpora; 2) spoken corpora; and 3) genre-specific spoken corpora.

1 **Spoken components of large general corpora**

We begin with spoken components of large general corpora as we have already referred to and accessed (to carry out search tasks) the spoken components of BNC and COCA. Here we add a little more detail:

**BNC Spoken Component**

Available at [http://corpus.byu.edu/bnc](http://corpus.byu.edu/bnc)

Description: The spoken part (10%) consists of orthographic transcriptions of unscripted informal conversations (recorded by volunteers selected from different age, region and social classes in a demographically balanced way) and spoken language collected in different contexts, ranging from formal business or government meetings to radio shows and phone-ins.

**COCA**

Available at [http://corpus.byu.edu/coca](http://corpus.byu.edu/coca)

Description: This is a large spoken component – around 90 million words at the time of writing – but it consists largely of broadcast data, e.g. interviews and talk shows, rather than informal, spontaneous speech. An interesting feature, however, is that the data is constantly updated, facilitating diachronic work (Lee 2010).

2 **Spoken corpora**

**The Limerick Corpus of Irish English (L-CIE)**

Available: This corpus is not publicly available.

Description (based on: [http://www.ivacs.mic.ul.ie/corpora/limerick-corpus-irish-english-l-cie/](http://www.ivacs.mic.ul.ie/corpora/limerick-corpus-irish-english-l-cie/)): Even though this corpus is not publicly available, it is worth considering its design features for two reasons:
1. It illustrates the range of criteria which have to be taken into consideration in designing a spoken corpus.

2. As it was designed to be comparable with CANCODE, we also get an insight into the composition of CANCODE: two for the price of one is always a good deal.

L-CIE is a one million word corpus of spoken Irish English with a design matrix derived from the influential and pioneering CANCODE and uses two axes for classification: context-type and interaction type.

The context-type axis has five categories: intimate, sociocultural, professional, transactional and pedagogic. While four of these categories are self-evident, the sociocultural category perhaps requires some explanation and it is explained thus on the website, ‘This category implies the voluntary interaction between speakers that seek each other’s company for the sake of interaction itself. The relationship between the speakers is usually marked by friendship and is thus not as close as that between speakers in the “intimate” category’.

The interaction type axis is explained thus: ‘… distinctions have also been made within the corpus between texts that were predominantly collaborative versus those that were non-collaborative, i.e. texts in which speakers give explanations and information or relate events and tell stories’.

The Santa Barbara Corpus of Spoken American English
Available at http://www.linguistics.ucsb.edu/research/santa-barbara-corpus#Intro

Description:

- It represents a wide variety of people of different regional origins, ages, occupations, genders, and ethnic and social backgrounds.
- It deals with mainly face-to-face conversation, but the corpus also documents many other ways that people use language in their everyday lives: telephone conversations, card games, food preparation, on-the-job talk, classroom lectures, sermons, story-telling, town hall meetings, tour-guide spiels, and more.
- About 250,000 words.

The Longman Corpus of Spoken American
Available: This corpus is not publicly available.
Description: http://www.pearsonlongman.com/dictionaries/pdfs/Spoken-American.pdf

- This corpus represents 12 regions of the USA.
- ‘Normal daily conversations’.
3 Genre-specific spoken corpora

The Switchboard Corpus
Available: https://www.ldc.upenn.edu/language-resources/data/obtaining
Description:

- A three million word corpus of spoken American English consisting of recorded telephone conversations from the early 1990s.

The Corpus of American Soap Operas
Available: http://corpus.byu.edu/soap/
Description:

- 100 million words of ten American soap operas from 2001 and 2012.


Spoken corpus research and lexis

Spoken word frequency

When considering the findings of spoken corpus research, word frequency is an obvious place to start: despite some of the difficulties of defining what a word is (see Chapter 3), it is relatively simple to calculate, and, as we shall see, observations about word frequency take us quickly and naturally into a consideration of discourse.

We begin with the commonplace observation that while the high frequency words in both spoken and written frequency lists are predominantly grammatical, the pronouns *I* and *you*, for example, are more frequent in spoken English than written English. As O’Keeffe, McCarthy and Carter (2007: 159) note, the higher frequency of these pronouns in conversation reflects the interactive nature of conversation. A further clear difference between writing and speech in terms of frequency is the fact that spoken language typically employs a narrower range of vocabulary than written language. This lexical density finding is reflected, for example, in the eye-catching finding by Biber and Conrad (2010) that 12 verbs account for almost 45% of occurrences of lexical verbs in conversation (the 12 verbs are *say, get, go, know, think, see, make, come, take, want, give, mean*).

Discussion

How far do you think the corpus finding above should influence course design and teaching?
Such *quantitative* research is clearly of interest to us from a teaching point of view, though, as we argued in Chapter 2, frequency alone never dictates what we teach. *Qualitative* analysis, however, adds value to these insights. O’Keeffe, McCarthy and Carter (2007), as we noted in the introduction, while drawing attention to the high frequency of the words *well*, *just* and *right* in conversation, go beyond this simple quantitative observation. They point out, through qualitative research, that while each of these words has a simple lexical meaning, they also have a discourse-marking function, and it is their discourse-marking function which contributes most to their frequency, e.g. Observe below how ‘J’ and ‘D’ use *well* and *just* in negotiating a potential disagreement over whether ELT methodology has progressed (TTT corpus):

J: And the director of studies was like this dragon called — and she was sitting in the back of the room sort of observing me and I wasn’t following the strict regime so I got hauled into her office, you know and told that, ‘This is what you’ve got to teach, this is how we learn’ and you know really put you know really put to book whatever the expression is.

D: Are we any better today than that though? Aren’t we a bit like her?

J: *Well*, I *just* …

D: *Well*, ‘we’, I don’t really know, you know, isn’t it a bit the same, you know you come along you do your TEFL course and you’ve gotta do it this way this way, this way and this way and then you go out and I’m sure directors of studies, cos there’s a bit of you know a power thing, isn’t there about all of that TEFL certification process that is self-perpetuating now because every there must be every director of studies in the world must have gone through a TEFL course by now and so what they look for is what they were told.

**Corpus Question**

What other devices do you see in the dialogue above which help the speakers to negotiate and mitigate disagreement?

See Appendix 1, Commentary 5

To continue with the theme of discourse markers, observe how I and H use *right* as a discourse marker below:

I: *Right*, Heather, tell us about meeting David, you were on the point …

H: Oh, *right*, well, I spent two months down in the south of France erm don’t know what to tell you and what not to tell you.
Discussion

Given the interactive importance of *well, just* and *right*, at what level do you think they should be taught, i.e. beginner, elementary etc.?

The frequency of the words *well, just* and *right* in their discourse-marking function in conversation reflects, as O’Keeffe, McCarthy and Carter argue, the interactive nature of conversation and underlines the importance of words which have ‘pragmatic functions in the organisation and management of conversation and in the speaker–listener relationship, particularly in terms of maintaining good relations’ (O’Keeffe, McCarthy and Carter 2007: 159). The effect of discourse concerns on spoken vocabulary frequency is further exemplified by Carter and McCarthy (2006) with reference to response tokens. Note, for example, how the response tokens are used by B in the two exchanges (from the TTT corpus) to show an interest in and a reaction to comments made by A.

A: Yeah, so we got married by special licence on the Friday and went to Japan, I think, two weeks later.
B: Amazing.
A: I kind of realised that I didn’t really wanna teach kids ever, cos.
B: Very wise.

*See Appendix 1, Commentary 6.*

Corpus Question

What other features of spoken language can you see in the two mini-dialogues above?

It is adjective uses, such as those exemplified above, that lead Carter and McCarthy (2006: 188) to observe:

Some adjectives and adverbs are many times more common in spoken language than in written language because of their frequent use as response tokens. These include: *absolutely, certainly, definitely, fine, great, good* … In spoken grammar, the term ‘response token’ better describes their function of referring to a whole preceding utterance rather than their word-class identity as adjective or adverbs …
As our example of response tokens shows, discourse concerns and interpersonal concerns can be closely related: response tokens, for example, oil the wheels of discourse but also create an affective bond by showing interest in what the speaker is saying, e.g.

J: Because I opened the paper and I saw an advert from Inlingua.
JB: Really!
J: …saying that they needed, you know, TEFL teachers, and er so I phoned them up and they said come for an interview so I flew to London cos in Jersey with mum and dad and they interviewed me and they said, ‘Right, you’ve got to go on a 15 hour training course’.

**Corpus Question**

What other features of spoken language do you see in the dialogue above, e.g. discourse markers?

See Appendix 1, Commentary 7

Biber *et al.* (1999: 516) point specifically to the influence of interpersonal concerns on spoken lexis, noting that the most common predicative adjectives in conversation are evaluative and emotive, e.g. *good*, *lovely* and *bad*. They also draw attention to the fact that some very frequent predicative adjectives (see examples below) can have a wide range of meanings, citing the example of *funny* which can be used to mean ‘amusing’ or ‘strange’ or even ‘unwell’, as in the examples below (Biber *et al.* 1999: 516):

- She sounds funny on the phone.
- I said I don’t think she’s funny. I think she’s idle.
- Funny though how people are like that.

The prevalence of such predicative adjectives in conversation reminds us of the centrality of evaluation to conversation (Pomerantz 1984; Aijmer 2005). It also reminds us that while discrete corpus findings are interesting, we are just as concerned with the general picture of language that emerges from the accumulation of these corpus findings.
Discussion

Keeping in mind the above discussion of the factors which influence spoken word frequency, how do you account for the following frequency statistics (based on the North American component of ICE) cited by McCarten and McCarthy (2010)?

- Actually is four times more common in speech than in writing.
- Absolutely and basically are in the top 1,000 words in conversation.
- Anyway is one of the top 300 words in conversation.

Frequency of collocations and lexical chunks

We have been at pains to stress that, while specific lexical frequency findings may be interesting, we need to interpret the cumulative significance of such research. Accordingly, the first thing we need to recall in this section is that multi-word units, however defined, play a huge role in both speech and writing. We discussed in Chapter 3, for example, how important collocations and lexical chunks are to fluent production and rapid comprehension. Further concrete evidence of the importance of collocations and chunks is provided when we consider the notion of merging word, collocation and lexical chunk frequency lists: if there were a unified frequency list of words, collocations and chunks, many collocations and chunks would be higher in the list than individual words (Shin and Nation 2008; O’Keeffe, McCarthy and Carter 2007).

More specifically in terms of frequency, Shin and Nation (2008: 339) conclude both that collocations are more frequent in speech than in writing and that ‘a large number of these [collocations] would qualify for inclusion in the most frequent 2,000 words of English, if no distinction was made between single words and collocations’. They point out, for example, that 84 collocations could be included in the level of the first 1,000 word types (Shin and Nation 2008: 345). When we consider spoken frequency lists of collocations or of lexical chunks the influence of interactive concerns on frequency, which we noted above in relation to individual words, is again apparent. Shin and Nation (2008), for example, observe that collocations such as you know, a bit and come on – first, third and twenty-first in the spoken collocation frequency ranking list respectively – reflect the interactional nature of conversation: you know is common in its role as a discourse marker; a bit functions commonly as a softening or hedging device as in, ‘I feel a bit disappointed’, rather than simply as a partitive, as in ‘a bit of peace and quiet’; come on is an informal imperative. None of them carries much propositional meaning. In light of the above, it is unsurprising that O’Keeffe, McCarthy and Carter (2007: 71) conclude that many of the most frequent chunks in CANCODE
fall into the pragmatic categories of ‘discourse marking, the preservation of face and the expression of politeness, acts of hedging¹ and purposeful vagueness …’.

Among the examples they provide of these pragmatic categories are the following (O’Keeffe, McCarthy and Carter 2007: 71–75):

**Discourse Marking**
- You know
- You know what I mean
- At the end of the day

**Face and politeness**
- I was going to say
- Do you think
- Do you want (me) (to)

**Hedging**
- Sort of
- A bit
- I don’t know

**Vagueness and approximation**
- And things like that
- And that sort of thing
- All the rest of it

**Corpus Question**
Identify in the short dialogue below (TTT corpus), examples of: spoken discourse markers; hedging.

H: I got a place at International House London, so we both did our course at the same time and then off I went so, you know, three week awful, awful, awful summer school in south of England, Hove, not Hove, somewhere awful.
D: It wasn’t EF was it?
H: No, it was even worse, I think and then I got, then yeah we decided, we had this grand master plan, me and Julie, we were gonna do a year in Paris, then we were gonna go for a year in Tokyo or Hong Kong, then we were gonna go for the third year in Rio de Janeiro and then we thought we’d you know think about settling down and being a bit more serious about life and so, anyway, that’s how I got into it. So, you know, we just packed
We have used vague language as a category of spoken language above and it is an area which has attracted research interest in recent years (e.g. Channell 1994; Cutting 2007). Channell (1994: 196), for example, stresses the pervasiveness of vague language in conversation:

Vague language forms a considerable part of language use. The corpora and other texts studied show many examples occurring in a wide range of contexts. This means we cannot, in any theory of language, treat it as the exception rather than the rule.

While such language is sometimes criticised on the grounds that it simply constitutes aimless ‘white noise’, or is used as a filler, Channell (1994: 197) argues that it is purposeful: ‘Vague expressions are not empty fillers, inserted by speakers to give processing time. They are deliberately chosen for their contribution to the communicative message’. The importance of chunks which carry out these pragmatic functions is underlined by Carter and McCarthy (2006: 73): ‘... the presence of... chunks plays an important role in the mutual protection of face and the smooth, sensible and sociable progression of the conversation’. If vague language needs further defence, we can turn to Lewis (2000) who argues that vague language is important in projecting one’s attitudes and personality and to O’Keeffe, McCarthy and Carter (2007) who maintain that vague category markers are of interpersonal value in indicating shared knowledge and understanding.

**Corpus Question**

The utterance below was made by an English language teacher talking about getting her first job in Portugal. What do you think she is referring to with the vague category marker ‘and stuff like that’?

- I didn’t want to quite settle down yet and stuff like that.
Discussion

How far do you think it is possible or desirable to teach vague language?

Spoken corpus research and grammar

Broadly speaking, I would argue that spoken grammar research is of potential relevance to ELT in two main ways:

1. It has provided new insights into grammatical phenomena which, while they have been part of the conventional ELT grammar syllabus, have been described only in relation to their written use.
2. It has shown that some non-canonical spoken grammatical features, not normally included in the conventional ELT grammar syllabus, are more systematic and pervasive than previously thought: such features, it has been argued, are of potential communicative value to the learner (McCarthy and Carter 1995).

In this section, we will consider ‘case studies’ of both types of grammatical feature outlined above:

Existential *there* and variable concord, e.g. there’s several reasons for beginning with this example.

We begin with existential *there* (there is/there are) as this structure is a staple of every elementary coursebook. The rule that *there is* is used with a singular noun phrase complement and *there are* is used with a plural noun phrase complement is very neat and teaching-friendly. Unfortunately, it turns out not to be the case, at least in spoken language: in the majority of cases, plural noun complements are used with *there’s* rather than *there are* (Biber et al. 1999; Cheshire 1999; Ruehlemann 2007). In other words, the default use of existential *there* in conversation is *there’s* for both singular and plural NP complements. Consider, for example, this illuminating dialogue between two children in Bolton in the late 1930s (Bolton Corpus):

S1: There’s no dragons today
S2: Yes, there is
S1: There’s not
S2: There is
S1: No, there’s not
S2: There is. Dragons eat people in Australia. There’s kangaroos there
S1: There’s not
S2: There are.
It is noteworthy here that, while there’s is clearly the default use for our young antagonists, S2 switches to the canonical there are form in the final utterance, perhaps in a last desperate attempt to sound authoritative about the important issue of the existence of dragons in Australia: a quintessential example of grammar as choice?

It is difficult to dismiss this use of there’s with a plural noun phrase complement as simply conversational carelessness given that it has an even longer history in the language than even the 1930s dialogue above suggests: Tagliamonte (1998: 106) argues that ‘…default singulars have existed in every century of the language’ citing an example from the sixteenth century:

There was many Dukes, Erles and Barons (c. 1533)

Indeed, Sobin (1997) goes so far as to argue that the use of there are for plural noun phrase complements is a ‘grammatical virus’ transplanted into the language by grammarians wanting to impose the grammar of Latin on the English language. Other commentators have accounted for the use of there’s for both singular and plural NP complements in slightly different ways: Ruehlemann (2007) suggests that the removal of the need to distinguish between singular and plural complements eases processing constraints; more specifically, Crawford (2005) argues that the default singular alleviates ‘… the cognitive difficulty of maintaining long turns …’. Cheshire (1999: 137), however, relates the default singular to turn-taking:

… existential there can be seen as a way for speakers to take the floor quickly and easily in lively conversation. Clearly it would be functional for such a useful construction to be shared and accessed as a prefabricated phrase…

It is worth noting that Cheshire (1999) considered this kind of concord variability to be on the increase: a case of ‘declining standards’ or ridding ourselves of a ‘virus’?

**Corpus Search**

Run a search on there’s at http://corpus.byu.edu/bnc/ (specifying the spoken component on the drop-down menu) and see how many plural noun complements you find with there’s.
Corpus Question

The multi-purpose like: What different meanings or functions can you attribute to like in the TTT extract below?

K: Well that’s the thing I mean when I was D.O.S-ing later in Dresden by then it wasn’t going to be my career it was another three years reflecting on life I found out I was like in the hierarchy if you like and it began to annoy me when people would turn literally kind of walk through the door and say Oh you know I’m passing through can I teach here for a few weeks Dresden/Prague/Vienna And I was like well NO you know this is a career I would hope you know this is something it’s like these people were just passing through but THEN that’s the problem you see we’re all passing through we’re here for six weeks six months six years ( )

See Appendix 1, Commentary 9

We saw in Chapter 4 that (be) like is one way of introducing reported speech: Carter and McCarthy (2006) argue that it is one of the more common uses of ‘like’. We will now consider its discourse-marking functions, beginning with clause-final uses of ‘like’, e.g.

Example 1:

There’s a very modern touch about the house – the way they’ve built it like.

Example 2:

S1: Good idea these revolving doors  
S2: Yes, keeps the draughts out  
S1: I always think of persons going round and round like when I come in  
S2: Yes, I often think of that, walking round and round like.

As both of these examples were taken from the Bolton Corpus (1937–1940), we can see that this structure, as was the case with there’s + plural NP complement, has history. Indeed, we should also note the observation by Miller and Weinert (1998: 307) that ‘… the discourse marker like is no late twentieth-century parvenu but has a recorded history in written English going back to the early nineteenth century’. While the longevity of the structure is interesting in itself, we need to know its function. Miller and Weinert (1998) refer to Hedevind’s (1967: 237) work on Yorkshire English in relation to utterance-final like, drawing attention to
his observation that this use was widespread in the North of England, appearing to function as a kind of down-toner. In relation to their own data, Miller and Weinert (1998: 331) conclude that ‘… declarative like-final clauses in the spontaneous conversation revealed that they can all be interpreted as countering potential inferences, objections or doubts’. We need to consider two other uses of like, exemplified below:

[These are attested examples, i.e. examples I have overheard and noted]

Example 1:

S1: What is IATEFL?
S2: It’s like an organisation.

Example 2:

S1: When I woke up, it was like … three o’clock.

Example 1 shows what Carter and McCarthy (2006) refer to as the ‘analogising function’ of like as the speaker avoids committing herself to an absolute definition, either because she genuinely doesn’t know what IATEFL is, or because she does not want her interlocutor(s) to feel she is an expert in the matter or that she feels she is an expert in the matter. Example 2, however, appears to show the way like marks a pause before newsworthy information (Carter and McCarthy 2006) as three o’clock is an unusual time to wake up. We can see, then, that like is a multifunctional discourse marker whose use needs to be discerned through plausible reconstructions of the speaker’s intention (which are not always easy to make).

Ellipsis

For examples of ellipsis, we return to our slightly surreal ‘revolving doors dialogue’ which reinforces McCarthy’s (1998) point that interesting features of spoken language are often present in the most banal of conversations.

S1: Good idea these revolving doors
S2: Yes, keeps the draughts out
S1: I always think of persons going round and round like when I come in
S2: Yes, I often think of that, walking round and round like.

There are two examples of ellipsis for which I have provided a plausible reconstruction below:

S1: [They are a] Good idea these revolving doors
S2: Yes, [they] keeps the draughts out.
Carter (1998: 45) explains that ellipsis typically involves the omission of the subject pronoun, copular or auxiliary and that, far from being some kind of aberration is ‘pervasive and endemic’ in conversational data and Carter and McCarthy (1995) argue that ellipsis is particularly common in lexical phrases, fixed and institutionalised expressions. To take but one example of an institutionalised expression, see you later omits the pronoun and auxiliary verb; however, if you included the pronoun and auxiliary verb, i.e. I’ll see you later, it might strike a quite different tone affectively (depending on stress and intonation). I would like to take an example from my own experience that shows the affective function of ellipsis, but this time from written rather than spoken language. While writing to a professional colleague from another institution, I was uncertain whether to end the letter:

‘Look forward to seeing you shortly.’

or

‘I look forward to seeing you shortly.’

I felt they struck quite different tones and I was not sure which to use as although I had friendly relations with this colleague she was in a formal relationship to the university (external examiner). Such grammatical choices are not a matter of life and death, but they can make a difference to how relationships are established and maintained.

Tails

We can return to two dialogues above for examples of tails (or ‘right dislocation’ as it is sometimes known), a structure we referred to in Chapter 4.

Example 1:

H: No, it was even worse, I think and then I got, then yeah we decided, we had this grand master plan, me and Julie, we were gonna do a year in Paris.

We see here the essential elements of the structure – the subject we has a co-referential element me and Julie placed after the conventional subject + verb + x structure.

In Example 1, the tail me and Julie could be explained as retrospective clarification on the grounds that H may feel the need to clarify the referents of we on picking up the story again.
Example 2:

It’s funny I would never have gone with somebody … desperately wanted to go on my own off somewhere. It’s funny that.

In Example 2, however, that does not clarify the subject it. Here we see the association of the structure with evaluation (e.g. McCarthy and Carter 1995; Timmis 2010).

We see further evidence of this evaluation function in the examples below from the Bolton Corpus, featuring first our ‘revolving doors’ conversationalists:

Example 1:

S1: Good idea these revolving doors
S2: Yes, keeps the draughts out.

Example 2:

She’s a good player, she is. (Bolton corpus)

While I would not wish to labour the point that non-canonical features are not necessarily new arrivals, this structure too has history and cannot be dismissed as merely a dialect form: Durham (2011) points to the use of tails in Victorian literature, while Lambrecht (2001) shows that tails occur across a wide range of different languages.

Spoken discourse

We observed earlier that one of the consistent findings of corpus research has shown that the borders between areas such as grammar, lexis, discourse and pragmatics, previously thought to be secure borders, are in fact permeable. In this section, we focus on how both spoken grammar and spoken lexis are shaped by discourse concerns.

Spoken discourse and grammar

The need to take a discourse perspective in order to properly account for spoken grammar is stressed by McCarthy and Carter (2001: 68):

The criterion we wish to press home here is that grammatical patterns exist across longer stretches of text, and that we must take a discoursal perspective that goes beyond the sentence or immediate utterance to establish the degree of overlap or otherwise in such patterns in written and spoken language.
We will discuss in this section a number of examples which show the influence of discourse concerns. We begin with tense and aspect choices as these often loom large in the minds of both teachers and learners and are normally prominent in the grammar syllabus. It is common in ELT, for example, to speak of ‘narrative tenses’, the assumption being that narratives are related through the judicious use of past simple, past continuous, past perfect. This, however, is not always the case: in oral narrative, for example, ‘speakers exercise considerable liberty in tense and aspect choice’ (McCarthy and Carter 2001: 60). A clear example of this is the use of the historic present in anecdotes and jokes: this cannot be some kind of performance aberration as it is a device routinely used in jokes; many jokes, for example, begin with the formula, ‘A man walks into a bar …’, e.g.

A man walks into a bar and turns to the people on the left-hand side of the bar and says, ‘You’re a bunch of idiots’ and to the people on the right-hand side of the bar, ‘You’re a bunch of fools’. A chap on the right-hand side gets angry and says, ‘Hey you, I’m no fool!’ The man says, ‘Well stand over there with the idiots then!’

I would argue that it is precisely because it is so natural to use historic present in this joke (which I rather like) that it passes unnoticed. We can also observe in this joke another feature of spoken language, the use of direct speech rather than reported speech to create a vivid impression. It is important to keep in mind, however, that this flexibility in the use of tense and aspect, as McCarthy and Carter (2001: 60) stress, is not random ‘but coincides with important segments of the narrative, where listeners are, as it were, taken in and out of the story-world in real time’.

A further example of the influence of discourse concerns on verb choice is provided by McCarthy (1998) discussing the use of used to and would for past narratives. Conventionally, ELT materials only make a straightforward grammatical distinction between these two forms: used to can be used for past habitual actions or states and would only for past actions. While this distinction is indeed true, it does not, however, fully account for how choices are made between the two forms in contexts where both are permissible for relating past habits. McCarthy (1998), however, takes a discourse perspective and points out that when used to discuss past habits, in oral anecdotes, for example, used to is typically used to set the context, while the speaker continues to relate the habitual actions with would, i.e. used to is used as a kind of preview, but details are related with would (McCarthy 1998) as the example below illustrates:

Thomas: Ach, everyone likes Irn Bru, but with vodka?
Leonard: I used to drink it that way at school. I’d steal some vodka from the drinks cabinet and then I’d mix it in an Irn Bru bottle.

(From a story called ‘Natural selection’ by Ian Rankin.)
While we have looked thus far at how grammar can be influenced by the need to create discourse, we need also to consider how spoken grammar can operate between speakers, i.e. interactively. We can return here to the use of *though* in conversation, briefly discussed in Chapter 4. Conrad (2004) makes two interesting points about the use of *though* in conversation: in contrast to writing, where *though* is often used initially or medially, in conversation *though* is often used clause-finally as a linking adverbial or interactively between speakers. Example 1 shows a clause-final use; Example 2 shows an interactive use:

Example 1:

I enjoy the job. I don’t know if I’ll be doing it in 10 years’ time *though*.

Example 2:

S1: It’s not nice  
S2: It’s funny *though*.

A further example of this kind of interactive grammar is provided by Tao and McCarthy (2001) who note how the relative pronoun *which* can be used turn-initially to complete the first speaker’s utterance. This kind of co-construction is illustrated below as S2 ‘completes’ the utterance begun by S1 so that it forms a co-constructed grammatical sentence. We should note, however, that S2 has not necessarily completed the sentence in the way S1 intended. I advise conversationalists to use the completion strategy sparingly as it can be irritating if over-used!

S1: The inspection is in November  
S2: *Which* is why we need to start the paperwork now.

We see two further examples of turn completion in the conversation below (TTT corpus):

S1: But there is I suppose the cult of beauty in America is whatever you do you must strive to be beautiful in a way that it’s part of the American Dream isn’t it the success the beauty (comes along) not that everybody’s got it [no] but S2: *Which* is strange cause all the American people I’ve met were singularly plain  
S1: Yeah yeah I remember being er in Indonesia as well and erm where I was teaching there was an American guy and er we got on very well but I went on a trip across to west Sumatra on a bus and I met some er in front of me there was a German and a Swede and we immediately had a kind of rapport [yes] that I didn’t have there’s some there was a certain cultural cultural background and we could understand each other in a way that with me and the American didn’t S2: *Which* is bizarre when you consider that most Americans are from a European heritage.
Spoken discourse and lexis

We turn our attention now to examples of how discourse concerns influence lexical choice and begin with how evaluation can shape discourse, recalling the important role evaluation plays in conversation as a whole. Pomerantz (1984) argues that, in conversation, when a speaker makes an assessment (i.e. an evaluative comment), it is typically followed by a second assessment rather than overtly signalled agreement (or disagreement). In the extract below, for example, K’s assessment that a new role is a ‘learning curve’ is followed by E’s evaluative comment that it is ‘a very steep learning curve’ rather than by an overt agreement marker such as ‘you’re right’ or ‘I agree with you’. We can also note that the second assessment is an upgrading or boosting of the first assessment, a typical phenomenon in such exchanges (Pomerantz 1984).

N: It’s another role a new role that you’re [everyone speaking at once here]
K: (  ) it’s a learning curve
E: It’s a very steep learning curve but erm.

We see a further example of the influence of evaluation on conversation in the observation by McCarthy that evaluative comments often occur before a change of speakers in the conversation:

Evison, McCarthy and O’Keeffe (2007) … present evidence for the way high-frequency evaluative adjectives such as lovely, awful, wonderful, funny trigger listener responses, with a strong preference for convergent response tokens. A considerable proportion of the occurrences of these adjectives occur immediately before speaker-change. The importance of these types of items is that they invite reciprocity and convergence and project seamlessly to the following utterance by the next speaker.

To remain with the theme of evaluation and the role it plays in conversation, we should note the argument (McCarthy 1998; O’Keeffe, McCarthy and Carter 2007) that idioms are often used evaluatively in the comment component of observation–comment patterns and are particularly common in the coda of anecdotes and stories (i.e. the part of the story which conveys the message or moral of the story). The patterns in which idioms are used indicate ‘the important interactive functions idioms can perform, creating and reinforcing interpersonal relations, projecting informality, camaraderie and social bonding’ (O’Keeffe, McCarthy and Carter 2007: 82).

Discourse sequences

Our discussion of though and our discussion of evaluative adjectives have taken us in the direction of how conversation is structured by discourse sequences. We
can briefly add further examples: Couper-Kuhlen and Thompson (2000), for example, outline what they call the Cardinal Concessive Schema, a schema which is designed to show how concession is prototypically managed in conversation. In the dialogue below, for example, S1 would have been within her rights to say to S2 (and may even have thought), ‘No, nothing like a dacha, much smaller’. In the real dialogue, however, we see the Cardinal Concessive Schema in action where the statement (x) by S1 is followed first by an apparent agreement (S2) from S1 and then by disagreement (y) by S1, i.e. S1 first appears to agree with S2 before indicating mild disagreement:

S1: … and they go off everybody has their little house generally most people have a little country house
S2: Like a dacha whatever you call it (x)
S1: Yeah (x1) /but small, really tiny (y) …

It must be stressed that Couper-Kuhlen and Thompson (2000) do not present the Cardinal Concessive Schema as an invariable sequence, nor argue that each part is always present: they argue that it is prototypical sequence towards which speakers generally orient.

It will be interesting to look next at direction-giving as a discourse sequence as direction-giving is a staple of ELT coursebooks. Scotton and Bernsten (1988: 378) focus on the typical structure of direction-giving and the features which characterise the sequence:

Naturally-occurring direction-giving typically has a number of parts: an opening sequence, the directions themselves, a pre-closing and possibly a closing. Further, the directions are frequently preceded by, or are interrupted with, orientation checkers and parenthetical comments. Confirmation checkers occur throughout the exchange.

**Corpus Question**

Identify the confirmation checkers in the two conversations below (Scollon and Bernsten).

*See Appendix 1, Commentary 10*

**Conversation 1**

S1: How do I get to the Vet Clinic?
S2: Oh, my God. It’s far, but, um. It’s on, it’s Shaw Lane is right, if you keep going up here you’ll hit Shaw Lane and it’s going that way. But go a little bit more and then cross the parking lot and it goes that way.
S1: Umhum.
S2: And then cross the road and just follow that sidewalk all the way down. If you keep following it you’ll see a sign that says Vet Clinic. But go all the way down there and cross the road and just follow the sidewalk down.
S1: Okay.
S2: Okay?
S1: Thanks a lot.

Conversation 2
S1: How do I get to the Vet Clinic?
S2: The Vet Clinic? Oh man, it’s a haul (pause) Uh, okay. Go up to the bridge, and that’s on Farm Lane. And you want to go right all the way up until you pass—um. Do you know where—um. Are you familiar with the campus?
S1: Pretty much.
S2: Do you know where the (pause) Uh Agriculture Hall is?
S1: Yeah, I know.
S1: Okay. When you see the Agriculture Hall, it will be on your right hand side. You want to go left again down towards Akers. Then the Vet Clinic should be on that road, and it says Vet Clinic.
S1: It does?
S2: Yeah. You can’t miss it. It should be just past the intersection of Bogue Street and the street you’ll be on (rising intonation).
S1: Okay.
S2: All right.
S1: Thank you very much.

Spoken corpus research and pragmatics

It is but a short step from discourse concerns to pragmatics (and frequently difficult to disentangle the two, as readers may have noticed), and we will consider now two studies which focus on the difference between coursebook treatment of pragmatic functions and corpus-attested realisations of these functions. We begin with Holmes’ (1988) study of the treatment of doubt and certainty in coursebooks. Holmes begins by asserting the importance for learners of being able to express doubt and certainty, in particular because the function of expressing doubt and certainty can overlap with politeness strategies. Holmes points out that doubt and certainty can be expressed not only through modal verbs (e.g. *may*, *might*, *could*) but also lexical devices (among other ways): lexical devices include words such
as *perhaps, maybe, unlikely* etc. She sets out the challenge for materials writers (Holmes 1988: 24) in preparing learners for this aspect of communicative competence:

Since they [materials writers] clearly cannot cover every possible way of expressing uncertainty and politeness, selection is necessary, and one can therefore ask what criteria are used in making the selection. Simplicity is one often cited criterion. Given, for example, that epistemic modality may be expressed prosodically, syntactically, lexically, and through discourse markers … it is undoubtedly true that lexical devices are the easiest of these to acquire.

While lexical modal devices may be the easiest modal devices to acquire, and frequently used by native speakers, Holmes (1988) points out that many coursebooks (though we need to note that Holmes was writing in 1988) devote a disproportionate amount of time to modal *verbs*. To take a concrete anecdotal example of this emphasis on modal verbs, I once set up in class a typical coursebook activity where the learners were supposed to work in groups and speculate about the famous mysterious disappearance of the crew and passengers from the ship *Marie Celeste*. The purpose of the activity was free practice of the past modal forms: *may/might/must/can’t have + past participle*, e.g. ‘They *might have fallen off* the ship’. One (very able) student, however, insisted on using only lexical devices such as *perhaps*, e.g. ‘Perhaps they fell off’. When challenged, he explained that the others wouldn’t understand him if he used the modal verb past forms. He had a point: I was making my learners run before they could walk: why then do modal verbs have such a prominent place in the ELT syllabus? One reason, as mentioned in Chapter 4, may be that modal verbs constitute what Thornbury (2000) has termed ‘Grammar McNuggets’: they are conveniently packaged into well-structured lessons (present–practise–produce – PPP) and easily consumed. Whether PPP lessons on modal verbs are nutritious in the long-term is open to doubt.

The focus of the study by Boxer and Pickering (1995) is on the speech act sequence of complaint/commiseration. As was the case with the Holmes study, they are interested in the differences between coursebook representation of the speech act and ‘real-life’ examples. Their specific focus is on indirect complaints which they define as ‘complaints where the addressee is not responsible for the cause of the complaint’ (Boxer and Pickering 1995: 45). They exemplify indirect complaints and commiseration thus (complaint in italics; commiseration in bold):

[Two female graduate students in a departmental library]

A: *They never have what you need in here. You’d think they’d at least have the important books and articles.*
B: They didn’t have what you were looking for?
A: No.
B: *That’s typical.*
The exchange of complaint and commiseration is, Boxer and Pickering (1995) argue, above all a social bonding strategy and, therefore, potentially important for learners who *are in a native speaking environment*. From their survey of seven ELT coursebooks Boxer and Pickering (1995: 52) conclude that (we must again note the date of the reference):

… many ELT texts that are currently popular for the teaching of functions continue to concentrate on the acquisition of linguistic competence, with insufficient attention to a fuller communicative competence. This is evident in their focus on mitigating or softening devices that make complaining less confrontational, and in the lack of contextual/interlocutor information that is necessary for the teaching of sociolinguistic competence.

Such ways of exchanging complaint and commiseration are, as Boxer and Pickering (1995) fully acknowledge, culturally determined, and may not be appropriate for learners outside a native speaking environment. This brings us to the important questions of the suitability of native speaker norms, linguistic, sociolinguistic or pragmatic, as a target for learners.

Discussion

How far do you think it is possible or desirable to teach complaint/commiseration routines such as those illustrated above?

Syllabus, materials and methodology

Having reviewed some of the findings of spoken corpora and discussed how these findings need to be seen in the light of questions about the ownership of English, now we need to take the discussion further into the classroom. In this section, we consider specific proposals for syllabus, materials and methodology for teaching spoken language. We also review the actual impact spoken language has had on materials.

Syllabus selection criteria

It can be tempting when reading spoken language research to assume that an area so intrinsically interesting and illuminating must have an application in ELT classrooms, or at least in ELT classrooms where the focus is on speaking. We must beware, however, of what Larsen-Freeman (2002) calls the ‘reflex fallacy’, i.e. the mistaken assumption that because something can be described, it has to be taught. This cautionary note applies to all syllabus decisions, not just to spoken language, as Timmis (2012: 518) argues: ‘… there is always a pedagogic filtering
process which mediates between the findings of language research (whether about spoken language or not) and practice’.

Some of the questions we ask of any potential item to be included remain relevant, e.g.:

1. Is the item useful?
2. Is the item frequent?
3. Is the item complex?

However, in the case of spoken language, Timmis (2012: 518) argues that another question becomes especially relevant:

4. Is the item socioculturally appropriate?

We will postpone discussion of this fourth question, however, to Chapter 8 when we take an overtly critical perspective on corpus research from a sociocultural standpoint.

An alternative way to address the spoken language syllabus issue is to ask the basic question:

What will the spoken language feature enable the learner to do communicatively?

Will it, for example, enable the learner to strike an appropriately informal tone? Will it enable the learner to buy time in the conversation? Will it enable the learner to respond appropriately and so on?

**The spoken lexis syllabus**

Frequency is a commonsense criterion for lexical selection and we now benefit, as we have seen, from word, spoken word and collocation frequency lists. As we established in Chapter 2, frequency can never be the sole criterion: if we applied the frequency criterion rigidly to spoken language then learners would learn little but grammar words (e.g. articles, prepositions, auxiliary verbs) in the early stages; they would also learn spoken discourse markers such as *well* before they could generate spoken discourse to bother marking. This rather flippant remark needs to be qualified by the serious observation that there *is* a case for teaching spoken discourse markers sooner than has traditionally been the case (O’Keeffe, McCarthy and Carter 2007). A further complication with frequency, of academic interest at least, is whether we should teach the metaphorical use of a word before its literal use in cases where the metaphorical use is more common, e.g. ‘My exam was a *nightmare*’.

Look at the screenshot below and consider the metaphorical uses of *nightmare* compared with the literal use.
Figure 5.1 BNC concordance lines: nightmare
All the other criteria we suggested that is needed to lexical selection apply to spoken lexis, i.e. we need to take into account learners’ interests, needs, level and cultural sensitivities, as well as pedagogically sensible ways of grouping words whose individual frequency may vary.

Timmis (2012) suggests that, in general, we should prioritise very common, socially unmarked spoken lexis, though it is open to question whether any language use can be totally unmarked socially, i.e. not identified with a particular social group. There are, however, some reasonably clear cases: cool and awesome are response tokens typical of the young, for example. By way of contrast, the grammatical frame ‘how did x go?’ does not seem to me to be particularly marked socially. While I have never seen this frame in a coursebook, I would argue that it is very useful as it can generate a number of polite but informal, everyday questions, e.g.:

- How did your exam go?
- How did the party go?
- How did your job interview go?

A focus on frames is also suggested by O’Keeffe, McCarthy and Carter (2007) who note how the frame a bit of a can be extended to include useful phrases to describe difficulties such as: a bit of a problem/hassle/performance/nuisance.

In relation to the selection of spoken lexical chunks, O’Keeffe, McCarthy and Carter (2007: 71) suggest that we organise our selection around pragmatic categories: ‘We would argue … that it is in pragmatic categories rather than in syntactic or semantic ones that we are likely to find the reasons why many of the strings are so recurrent, and in the idea of chunks as frames that we will find the most pedagogically useful “handle” on chunks for vocabulary teaching and learning’. They suggest a number of pragmatic categories which, as they note, refer to the speaker–listener world rather than the content or propositional world (O’Keeffe, McCarthy and Carter 2007):

- discourse marking,
- the preservation of face and the expression of politeness,
- acts of hedging and
- purposive vagueness.

The spoken grammar syllabus

Arguments for a spoken grammar syllabus have generally been made in relation to potential benefits for affect and/or fluency. In making the case for at least raising learners’ awareness of features such as heads, tails and ellipsis, McCarthy and Carter (1995) stressed, (and have consistently stressed since then) the work that such features do in terms of establishing and maintaining relationships: they form part of an interpersonal grammar rather than a transactional grammar. We
have already seen how tails tend to convey evaluation and, it has been argued, they invite the listener to share that evaluation. To return briefly to the example of reported speech in conversation (discussed above), McCarthy (1998) points out that in conversation it can, and often is introduced by ‘I was saying’, i.e. the reporting verb is in the past continuous. This structure serves the function of summarising the gist of a conversation rather than the exact words (McCarthy 1998) and, in my experience is often used to summarise a conversation for the benefit of a newly arrived participant. This structure does not, however, to my knowledge, appear in any mainstream coursebooks. McCarthy and Carter (1995) also stress how certain structures such as ‘heads’ are listener-sensitive. Consider this example from my own experience:

S1: My dad has always wanted to be a postman  
S2: My friend Phil, his wife, she did that.

S2’s utterance certainly violates canonical grammar with its ‘double subject’ (his wife + she) and its non-use of the canonical genitive (my friend Phil’s wife), but the utterance can be regarded as listener-sensitive in the way it stages the information for S1 (who didn’t know Phil or his wife).

For Mumford (2009: 140–141), however, certain native speaker spoken grammar features need to be included in the spoken language syllabus not simply for affective purposes but to promote fluency. He lists the following as examples of native spoken grammar features, suggesting that such features could help learners to be more fluent:

• Native speakers often do not discriminate between there is and there are or less and fewer.
• Native speakers often use elliptical forms.
• Native speakers often use declaratives as questions.
• Native speakers are flexible in their word order, e.g. through the use of headers, tails and fronting.
• Native speakers use pausing devices such as ‘er’ and ‘um’.
• Native speaker language is repetitive and native speakers use chunks such as ‘I see’, ‘I think’, ‘you know’, ‘kind of’ to increase fluency.

What needs to be emphasised, however, is that native speakers do not use such features randomly. There is more to spoken language than forms.

Discussion

How far do you agree that the features above can or should be taught?
Syllabus and sequence

Even if consensus could be reached on the spoken lexical and grammatical items to be included in a syllabus, there would remain the question of how such a syllabus would be sequenced and instantiated. One possibility is to adopt a text-based approach as proposed by Willis (2003) and Tomlinson (2013), though they did not propose such an approach specifically in relation to teaching spoken language. A text-based approach to teaching spoken language is, however, proposed by Timmis (2010). This approach involves what we might call ‘informed opportunism’: the selection of items to teach is informed by the discussion we have had above about likely candidates for inclusion in the syllabus. This approach is opportunistic insofar as actual decisions about which features should be highlighted are deferred until the spoken texts in a given set of materials have been reviewed. There are, I would argue, two main advantages to adopting such an approach to teaching features of spoken language:

1. If the spoken texts which are a normal part of the materials or course, or selected for the course, are authentic or at least well-constructed (see below), then spoken language features will be experienced in their natural discourse context before becoming the specific focus of attention.

2. Teachers are probably in the best position to apply the selection criteria we have discussed above, particularly in relation to sociocultural concerns, as they will probably be better informed about the needs, goals and attitudes of the learners in their context.

An example of applying such an approach to language selection in practice is provided by Timmis (2010), based on the TTT corpus. Operating the criteria we discussed above, he considered the following in the table below from these conversations to be candidates for teaching with a multilingual group of adult learners in the UK. It is important to note here that Timmis (2010) considered this to be a provisional framework in relation to a specific group and that one would only arrive at a more carefully crafted syllabus through trying out the syllabus in the classroom and by examining far more data. Developing a spoken language syllabus is, then, likely to be an iterative process involving a certain amount of trial and error, but we need to start somewhere. The example below of a starting point based on the small TTT corpus is offered in all humility.
Table 5.1  Spoken language ‘syllabus’

A  LANGUAGE

1  Colloquial lexis

<table>
<thead>
<tr>
<th>Feature</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>General colloquial vocabulary</td>
<td>but I didn’t get picked to go there so I ended up in Lisbon</td>
</tr>
<tr>
<td>Colloquial phrasal verbs</td>
<td>You pick up certain things don’t you</td>
</tr>
<tr>
<td>Delexical verbs</td>
<td>so we had a little dance</td>
</tr>
<tr>
<td>Colloquial lexical chunks</td>
<td>And it was a case of oh right I could go here or there or here or there and I’ve never been to Rome ‘Yes, let’s go there …’</td>
</tr>
<tr>
<td>Colloquial collocations</td>
<td>Dead simple</td>
</tr>
<tr>
<td>Colloquial idiom/metaphor</td>
<td>IFS must be a big eye-opener for you</td>
</tr>
<tr>
<td>Lexical creativity</td>
<td>I think you’re probably busier admin-wise than I am</td>
</tr>
</tbody>
</table>

2  Colloquial grammar

<table>
<thead>
<tr>
<th>Feature</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spoken/written word order differences</td>
<td>… but they were quite a decent school to work for though weren’t they?</td>
</tr>
<tr>
<td>Tails</td>
<td>Nice, these olives</td>
</tr>
<tr>
<td>Reported/direct speech</td>
<td>They’d go ‘hey come here and have a beer!’</td>
</tr>
<tr>
<td>Ellipsis(^2)</td>
<td>… got married to get a job? Oh I’m shocked</td>
</tr>
</tbody>
</table>

B  INTERACTION STRATEGIES

1  Conversation strategies

<table>
<thead>
<tr>
<th>Feature</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic management</td>
<td>Going back to madly in love, wasn’t there some story about getting married to get a job?</td>
</tr>
<tr>
<td>Repair strategies</td>
<td>It’s quite a common thing isn’t it to take a year, what do they call them, they have an official year off from their work</td>
</tr>
<tr>
<td>Clarification strategies</td>
<td>What do you mean by ‘planned’?</td>
</tr>
<tr>
<td>Discourse markers</td>
<td>and then we thought we’d you know think about settling down and being a bit more serious about life and so, anyway, that’s how I got into it</td>
</tr>
</tbody>
</table>
2 Good listenership

<table>
<thead>
<tr>
<th>Feature</th>
<th>Example</th>
</tr>
</thead>
</table>
| Response tokens          | A: Yeah, so we got married by special licence on the Friday and went to Japan, I think, two weeks later  
B: Amazing                |
| Response chunks          | A: I kind of realised that I didn’t really wanna teach kids ever, cos  
B: Very wise               |
| Backchannelling          | A: I wouldn’t have dared to go by myself, I think it was only because there were two of us  
B: Really?                 |

3 Vague language

<table>
<thead>
<tr>
<th>Feature</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placeholder words</td>
<td>So much of teaching is a personality thing though</td>
</tr>
<tr>
<td>Vague category markers$^3$</td>
<td>I didn’t want to quite settle down yet and stuff like that</td>
</tr>
<tr>
<td>Hedging$^4$</td>
<td>Well, it’s I think it’s something that you fall into a bit though, isn’t it?</td>
</tr>
</tbody>
</table>

Notes
1 ‘This category [delexical verbs] embraces extremely high frequency verbs such as do, make, take and get … in their collocations with nouns, prepositional phrases and particles. They are termed “delexical” because of their low lexical content and the fact that their meanings in context are conditioned by the words they co-occur with …’ (O’Keeffe, McCarthy and Carter 2007: 37–38).
2 The absence of elements which would be required in a canonical written sentence.
3 Phrases which allow ‘speakers to refer to categories of people and things in an open-ended way which calls on shared cultural and real-world knowledge to fill in the category members referred to only obliquely’ (O’Keeffe, McCarthy and Carter 2007: 74).
4 Devices which allow speakers to ‘modify utterances to make them less assertive and less open to challenge or rebuttal’ (O’Keeffe, McCarthy and Carter 2007: 74–75).

Spoken language and methodology

Discussion of the possibility of a text-based approach to spoken language selection brings us naturally to a discussion of methodology. While methodology is not the primary focus of this book, we do need to consider whether syllabus recommendations from corpus research can actually be implemented successfully in the classroom, particularly when such recommendations present a challenge to orthodoxy. In relation to a text-based approach for spoken language, Timmis (2005) follows Willis (2003) in recommending that texts are processed for meaning before learners focus on specific aspects of language. We review here some of the methodological proposals made for dealing with spoken language in the classroom. An early and influential proposal was made by McCarthy and Carter (1995) who suggested an I-I-I framework (Illustration–Induction–Interaction): illustration through examples in context; induction through guided discovery; and interaction through discussion of the use of the target feature. An
addition to this framework was suggested by Timmis (2005) who proposed the use of tasks where learners compare their expectations of language use with the reality of language use, e.g.:

1. Look at the Chris/Trudy/Enid dialogue below and put brackets round any words you think might not be necessary in informal spoken English, e.g.

   How are you?
   I am fine thanks and how are you?
   I am fine too. It’s very cold isn’t it?
   How are you?
   [I am] fine thanks and [how are] you?
   [I am] fine [too]. [It’s] very cold isn’t it?

2. Remove words from the conversation below which you think could be omitted.

   Chris: Oh, hello Enid. It’s nice to meet you.
   Trudy: Hello, how’s the family?
   Enid: They’re very well. How are your family?
   Trudy: They are smashing.
   Chris: It’s a beautiful day, isn’t it?
   Enid: Yes, it is.
   Chris: Is your business booming?
   Enid: Well, it’s not too bad you know.

3. Now listen/watch and check which words are missed out on the video dialogue.

The emphasis of both these frameworks (McCarthy and Carter 1995; Timmis 2005) was strongly in favour of awareness-raising with little or no emphasis on production. The case for a productive component in such a framework, however, has been made by Mauranen (2006) and Jones (2007), though in slightly different terms: Mauranen (2006) suggests that production is necessary for acquisition while Jones (2007) argues that production can lead to noticing as learners consider how to use a particular feature. It is interesting to note here that in a small trial of his framework, Timmis (2005) observed that many learners actually wanted a productive component. In contrast to these proposals, which focus predominantly on awareness-raising, however, McCarten and McCarthy (2010) argue that precisely because the nature of the target language is different, the methodology should remain familiar, i.e. PPP (as indeed is the case in the Touchstone series).
Discussion

Do you feel that tasks focusing on spoken language should be ‘noticing only’ or should they have a production component?

In their review of the treatment of spoken language in coursebooks, Cullen and Kuo (2007: 379) point out that when spoken language is dealt with in coursebooks, the methodology appears to have been influenced by the McCarthy and Carter (1995) I-I-I frameworks and perhaps also by Timmis’ (2005) framework:

After a task to check global comprehension of the text, the learners’ attention is drawn to the target feature of spoken grammar… and its communicative purpose and use in the listening text is either explained or explored through some – usually rather brief – questions for discussion. There is then typically a short practice activity, where the learners are required to use the feature in a fairly controlled setting.

To reinforce the criteria for spoken language tasks, it will perhaps be useful to look at a suggestion for an activity that, as it stands, does not, in my view, meet the criteria we have discussed. The instructions for the activity are below:


- Say a short sentence, e.g. ‘John lost his wallet’, and then toss a coin.
- Ask the students to change the sentence with either a header or a tail while the coin is in the air. When the coin lands, tell the students whether it was heads or tails.
- If heads, students who used a header are the winners and vice versa.
- Students play in threes, one to toss and say a sentence in its conventional form (these can be given on a piece of paper if necessary) and two callers, who change the sentence with either a header or tail.

The activity appears to me to be problematic as the sentences provided by the teacher are random and, crucially, decontextualised – this means that the crucial issue of social appropriacy in the use of spoken grammar cannot be considered. In other words, the spoken forms in this task (at least as seen in isolation) appear to be presented as neutral alternatives to the canonical form without regard to the contexts and purposes for which they are appropriate. One could add that points seem to be scored as much by chance as by skill – speaking as a naturally competitive person, this seems to me unsatisfactory.
Spoken language and materials

We have already touched briefly on one specific proposal for classroom materials, Timmis’ (2010) suggestion that teachers’ personal anecdotes might provide one kind of material. However, we clearly need alternative options. An important question to address is the issue of authenticity: Gilmore (2004: 367) makes a case for authenticity:

If our goal in the language classroom is to prepare learners for independent language use, then surely we are obliged at some stage to present them with realistic models of discourse, messy and unpredictable as it is. So far I feel that just how much reality learners can cope with at different stages in their learning has not been adequately addressed.

At the same time, Gilmore (2004) notes some of the ways in which coursebook dialogues typically differ from authentic dialogues:

- length
- turn-taking patterns
- lexical density
- number of false starts, repetitions and pausing
- frequency of terminal overlap or latching
- use of hesitation devices or back-channelling.

Similar criticisms are made by Walsh (2010: 334):

It is uncontroversial to say that, at the time of writing, the following common features of everyday interaction are often missing from invented textbook dialogues:

- There are few response tokens: speakers do not generally acknowledge each other’s contributions.
- There are few if any repetitions, false starts or hesitations.
- The language used is often overwhelmingly transactional; there is often scant attention to the relationship between the speakers.
- There are often very few signs that the dialogue is presented as a ‘collaborative enterprise’, where both speakers (or several) are working together to achieve a common goal.

While authentic dialogues will probably have the important advantage of providing a natural context for the features we are interested in, they potentially present difficulties for learners, level of language being the most obvious potential problem. Among further potential drawbacks of authentic dialogues for classroom use, McCarten and McCarthy (2010) note the following:
• The use of puns or jokes (which are often culture-bound).
• The presence of dialect, colloquial or non-standard forms.
• The difficulty of identifying a conversation boundary.

We can add to these potential problems, Cook’s (1998: 61) observation that much everyday conversation is ‘inarticulate, impoverished, and inexpressive’, though one might qualify this comment by saying that this is how everyday conversations might seem to the outside observer rather than the interlocutors themselves. A further challenge with authentic dialogues is identifying what might be considered a typical conversation, as Dellar (2004) observes:

… what corpora can’t show us is what typical conversations look like. It’s not possible, for instance, to access ten typical conversations had by people talking about what they did last night or to look at the 20 most common ways of answering the question ‘So what do you do for a living, then?’ As such, if we want to present our students with models of the kinds of conversations they themselves might actually want to have, we are forced to fall back on our (actually ample) experience of such conversations in order to script them.

We should perhaps add here, that it is just as difficult to construct a typical conversation as it is to find one in a corpus. We may be able to construct a conversation full of typical spoken language features, but this does not mean that the conversation will be typical in terms of purposes, content, relationships, setting and so on.

With the above considerations in mind, McCarten and McCarthy (2010: 20) provide specific guidelines for the construction of what we might call ‘corpus-informed dialogues’. Among these guidelines are the following suggestions:

• Keep turns generally short, except for narratives. Where one speaker ‘holds the floor’ build in listener back-channelling and non-minimal responses.
• Allow speakers to react to the previous speaker.
• Don’t overload speech with densely packed information; ensure a balance of transactional and relational language and an appropriate lexical density.
• Include some repetition, rephrasing, fragmented sentences and other features of speech, but maintain transparency.
• Keep speakers ‘polite’ and not confrontational or face-threatening.

Ultimately, the question of using authentic or scripted, corpus-informed dialogues will depend on our learners’ level, needs and wants, but there is no need, I would argue to commit totally to one or the other. There is a need, however, for an awareness of the potential advantages of both authentic and scripted dialogues.
Conclusion

In conclusion, we have reviewed in this chapter findings from spoken corpora of potential relevance to ELT, noting that many of these findings can challenge traditional grammatical descriptions which have held sway. We have considered arguments that grammatical choice is significantly influenced by affective and discourse considerations, acknowledging that this kind of research can be deeply unsettling for those with firmly-held views of what grammar is, how it works, and how it should be applied. We have also touched upon the sociocultural implications of spoken corpus research which feature in Chapter 8. Finally, we reviewed some of the possibilities for syllabus, materials and methodology in the light of both the findings of spoken corpus research and the accompanying sociocultural debate. The importance of spoken language research in relation to contemporary language teaching is summed up by McCarthy and Carter (2001: 53) thus:

… language pedagogy that claims to support the teaching and learning of speaking skills does itself a disservice if it ignores what we know about the spoken language. Whatever else may be the result of imaginative methodologies for eliciting spoken language in the second-language classroom, there can be little hope for a natural spoken output on the part of language learners if the input is stubbornly rooted in models that owe their origin and shape to the written language.

Corpus Question

Identify at least five features typical of spoken language in the TTT dialogue below:

J: yeah, so then I just looked in the newspaper, thought, “Well shall I just go into primary school teaching or…” then I saw this job interview in Portugal so, just decided to go to the interview and went and ended up going down to this hotel in London into a bedroom being interviewed and no qualifications and nothing at all and they offered me like free Portuguese lessons and free flight and free this that and the other, free accommodation so I just thought, “Well is it primary school teaching or is it jump on a plane and head off to Lisbon” so that’s why I did it and also I mean didn’t ever think I’d sort of end up with it, you know, as a career or anything at all

I: But they were quite a decent school to work for though weren’t they, I’m surprised…

See Appendix 1, Commentary 11
Discussion

How desirable it would be to teach the features you have identified?

References


Chapter 6

Corpora and the classroom

Introduction

Thus far we have largely been concerned with the potential of descriptive findings from large corpora such as the BNC to inform our teaching and syllabus or materials design. In this chapter, while these perspectives remain relevant, we turn our attention towards the pedagogically-motivated design and exploitation of corpora. Specifically, we focus on three areas in which corpus research and corpus data has moved closer to or into the classroom:

1. The development of corpora of English used by learners (learner corpora).
2. The development of teaching-oriented corpora.
3. The direct use of corpus data by learners, usually known by Johns’ (1991) term as ‘data-driven learning’ (DDL).

While we have stressed in previous chapters how teachers and materials writers can gain from critical engagement with corpus research, in this chapter we will be more concerned with how practitioners can develop and exploit corpora for classroom purposes.

Learner corpora

What is a learner corpus?

Learner corpus research, according to Granger (2002: 3), has ‘created an important link between the two previously disparate fields of corpus linguistics and foreign/second language research’. It is an area of obvious interest to us, then, particularly as Granger (2002) stresses that learner corpora contribute not only to research but also to improving second language teaching and learning. Leech (1981, cited in Pravec 2002: 81) provides a straightforward definition of a learner corpus: ‘A learner corpus is a computerized textual database of the language produced by foreign language learners’. We can add to this definition that learner corpora may be ‘monolingual’, i.e. the data is gathered from speakers of one language, or
‘multilingual’, i.e. the data is gathered from speakers of several different languages. While in theory learner corpora may include data gathered from participants in various settings, in practice many learner corpora are based on data from classroom contexts, e.g. learners’ essays or discussion tasks in the classroom. We should also note that, while more learner spoken corpora are beginning to emerge, the time and investment needed to compile them means that they are far less common than learner written corpora (de Cock 2010). In this section, we will consider the design criteria for learner corpora with some case studies of well known, large-scale learner corpora. We will also discuss an indicative section of findings for learner corpus research and the potential practical applications of this kind of data.

**Design criteria for a learner corpus**

Assembling a learner corpus, Pravec (2002: 90) notes, is quite a complex and time-consuming process. However, we should not be discouraged: while it is useful to know the optimal design criteria in order to assess the value of learner corpus data, the ‘quick and dirty option’ we discussed earlier in relation to corpus compilation is open to us in this domain too (see Bennett 2010 and Walsh 2010 for discussion of small scale class-specific corpora). The first consideration, Granger (1994: 26) argues, given ‘the very heterogeneous variety’ of learner English, is ‘a very precise description of the population represented in the corpus’. Drawing on Granger (1994: 26) and Tono (2003: 800) we can see that there are many variables which need to be considered:

- learning environment
- age
- proficiency level of the learners
- mother tongue
- stage of learning
- nature of the task which was the stimulus for the data, e.g. topic, genre
- setting, e.g. classroom or homework
- use of reference resources, e.g. dictionaries, grammar books.

In relation to the mother tongue criterion, the basic choice is between what de Cock (2010) terms Mono-L1 and Multi-L1 learner corpora. The JEFLL (Japanese EFL Learner) Corpus is an example of a Mono-L1 learner corpus (http://jefll.corpuscobo.net/).

The JEFLL Corpus is a collection of free compositions written by more than 10,000 Japanese-speaking learners of English. The corpus size is approximately 700,000 words. It consists of the subjects ranging from novice to intermediate levels, covering mainly junior and senior high school students in Japan. The essay task is carefully controlled so that each sub-corpus can be comparable across topics, proficiency, school years, school types, among others.
The International Corpus of Learner English (see page 000) is an example of a Multi-L1 learner corpus.

In the case of learner spoken corpora, de Cock (2010) points out that the most common task types used to generate data are:

- interviews
- discussions
- role plays
- picture descriptions
- prompted oral narratives.

All of these spoken task types will normally involve interaction with other learners and/or with an interviewer.

The importance of all this kind of background information about task types and learner data (e.g. nationality, age, level), Pravec (2002) points out, is that we can extract sub-corpora from the main learner corpus we are working with for a specific line of research. We might, for example, want to focus our research on a specific level, a specific nationality, a specific task or some combination of such factors. With learner spoken corpora, in order to facilitate comparison, we might also want to build a parallel ‘control’ corpus of native speakers performing the same kind of task.

Once the background variables have been specified and the data collected, the choice is then between ‘a raw corpus, i.e. a corpus of plain texts with no extra features added, or of an annotated corpus, i.e. a corpus enriched with linguistic or textual information, such as grammatical categories or syntactic structures’ (Granger 2002: 7). If the decision is taken to annotate the corpus, it should ideally be annotated with standard tagging software (e.g. CLAWS, discussed in the introduction) to facilitate comparison with native speaker corpora (Granger 2002). We need to keep in mind, however, as noted by Gilquin, Granger and Paquot (2007), that the non-standard nature of some learner data may affect the reliability of standard tagging software. If we are dealing with learner corpora, it is highly likely that we will be interested in error tagging: Bennett (2010: 79–81) provides a useful example of straightforward error tagging code and an example of a tagged text based on this code. The following sample error tags from the Cambridge Learner Corpus are provided by Hawkins and Buttery (2010: 7):

RN Replace noun, e.g. *Have a good travel* (journey)
RV Replace verb, e.g. *I existed last weekend in London* (spent)
MD Missing determiner, e.g. *I spoke to President* (the)
*I have car* (a)
AGN Noun agreement error, e.g. *One of my friend* (friends)
AGV Verb agreement error, e.g. *The three birds is singing*
The advantage of comprehensive error tagging of a learner corpus, Granger (2002) argues is that it may bring to light frequent error types of which researchers were not previously aware.

To reinforce how learner corpora are designed, and to preview the kind of research carried out with learner corpora, it will be useful to consider in some detail four specific corpora: the ICLE corpus (International Corpus of Learner English); the Cambridge English Profile Corpus; the NICT Japanese Learner English (JLE) Corpus; and the Louvain International Database of Spoken English Interlanguage (LINDSEI).

**ICLE (International Corpus of Learner English)**


The International Corpus of Learner English, a Multi-L1 corpus compiled under the auspices of the University of Louvain, can be regarded as a pioneer in the field. A basic description is provided on the website http://www.uclouvain.be/en-cecl-icle.html

This corpus contains argumentative essays written by higher intermediate to advanced learners of English from several mother tongue backgrounds (Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Tswana and Turkish). The corpus is the result of collaboration with a wide range of partner universities internationally. The first version was published on CD-ROM in 2002, and an expanded version, ICLEv2, featuring a built-in concordancer, was published in 2009. The corpus is highly homogeneous as all partners have adopted the same corpus collection guidelines.

The ICLE corpus was compiled with two main aims in mind (Pravec 2002): to investigate errors in the corpus and determine whether they are language-specific or universal; to investigate what makes learner writing sound ‘foreign’, e.g. the overuse or underuse of certain words or phrases compared with native speaker English.

**English Profile**

www.englishprofile.org

English profile is a multi-L1 written and spoken corpus compiled under the auspices of Cambridge University Press. The summary below is abridged from the description on the website:

The Cambridge English Profile Corpus (CEPC) is a corpus of learner English produced by students all over the world. … English Profile aims to collect 10 million words of data, covering both spoken (20%) and written (80%) language. Both General English and English for Specific Purposes are included. …The corpus covers levels A1-C2, and attempts to maintain a balance across a number of variables, including:

- educational contexts (e.g. primary or secondary, monolingual or bi-lingual)
• linguistic function (informative, suasive, attitudinal, socialising, and structuring discourse. These categories are taken from the T-series and can also be found in Appendix B of Antony Green’s book *Language Functions Revisited*.)
• type of interaction, e.g. casual conversation, formal presentation, oral exam, classroom discourse, role play etc. (spoken data only)
• first language of learners
• age range of learners, and other demographic information
• CEF level.

We can see, then, that the basic aim of the English Profile Corpus is to add flesh to the bones of the level descriptors of the Common European Framework (CEFR). The original level descriptors of the CEFR were expressed as can-do statements, i.e. a functional approach to describing the levels of the framework. English Profile research seeks to describe the grammatical and lexical resources typically available to learners at a given level. The grammatical and lexical features which seem to define a given level, i.e. they are shared by a large number of learners at that level, are termed ‘criterial features’ (Hawkins and Buttery 2010). Among the aims of English Profile research relevant to our concerns are the following questions:

• How do the different kinds of criterial features (lexical semantic, syntactic, discourse etc.) cluster together to define learner profiles in English? Which linguistic features realise which language functions across the CEFR levels?
• How does the profile of the learner vary depending on their L1? What are the pedagogical implications of such L1 effects for the learning, teaching and assessment of English?
• Which criterial features can be used as diagnostics of proficiency at the individual learner level?

**The NICT Japanese Learner English (JLE) Corpus**
http://alaginrc.nict.go.jp/nict_jle/index_E.html

The JLE Corpus is a large Mono-L1 spoken corpus including data from learners at different levels of proficiency. The description below is abridged from the website:

The source of the corpus data for the NICT JLE Corpus is the transcripts of the audio-recorded speech samples (1,281 samples, 1.2 million words, 300 hours in total) of the English oral proficiency interview test. The advantage of using the test data as a source is that each speaker’s data includes his or her proficiency level (nine levels) based on their test scores, which makes it possible to easily analyse and compare the characteristics of interlanguage of each developmental stage. This is one of the strengths of *The NICT JLE Corpus* that cannot be often found in other learner corpora.
We have also compiled a sub-corpus for comparison. It is a native English speakers’ corpus. This sub-corpus is considered to be quite useful for comparing the utterances of native speakers and Japanese learners. We were able to make this comparison by collecting the speech data of native speakers, conducting a similar type of interview to that of the SST.

A particularly interesting feature of the JLE Corpus is the parallel control corpus of native speaker data for comparison purposes. As the corpus is at nine different proficiency levels, it can also facilitate longitudinal studies.

LINDSEI (Louvain International Database of Spoken English Interlanguage)
The description below is abridged from the LINDSEI website:

The aim of this project was to provide a spoken counterpart to ICLE, containing oral data produced by advanced learners of English from several mother tongue backgrounds.

All 11 components follow the same structure, with c. 50 interviews made up of three tasks: set topic, free discussion and picture description. The interviews are transcribed and marked-up according to the same conventions, and each of them is linked to a profile which contains information about the learner, the interviewer and the interview itself. This information makes it possible to study the possible influence of certain factors on learner language. Note that, alongside the non-native varieties of English, a comparable corpus of interviews with native speakers of English has been compiled (LOCNEC), so that interlanguage and native language can be compared and the universal and L1-specific features of oral interlanguage identified. The fact that LINDSEI and ICLE have been built according to similar principles and share as many as ten mother tongue backgrounds also means that they can be used in combination with each other to compare spoken and written interlanguage.

It is worth noting here that the corpus can be used in conjunction with a comparable native speaker corpus (LOCNEC) and a comparable learner written corpus (ICLE). LINDSEI does not, however, capture natural conversation: there is a degree of artifice in the tasks which generate the data.

Discussion

Which of the four learner corpora above are most relevant to your teaching/research interests and why?
The purpose of this review of four learner corpora has been to show the various factors involved in learner corpus design and to give some flavour of the kind of research which can be carried out with such corpora. It is important to reiterate, however, that for our own teaching purposes it is quite legitimate to build smaller, less elaborate corpora. Indeed, there may be advantages in doing so: Mendikoetxea, Bielsa and Rollinson (2010), for example, show how the INTELeNG corpus, consisting of 80 essays by first year Spanish university students of English, was used to inform classroom materials. They used the data from their corpus to design grammar materials. These materials generally begin with a typical grammatical error taken from the corpus data, accompanied by the correct form and a grammatical explanation. Follow-up practice is then provided through a variety of exercises.

**Learner corpus research: Theory and practice**

From the descriptions of ICLE, the Cambridge English Profile Corpus, and the Japanese Learner English Corpus above, we can summarise the main aims of learner corpora:

- To identify differences between the English of learners who share the same L1 and native speaker English. This kind of research identifies typical L1-specific errors and items which learners overuse or underuse in comparison with native speakers.
- To compare the English of learners who speak different first languages with native speaker English. This kind of research can not only reveal typical L1-specific errors and underuse or overuse of given features but can also identify what seem to be universal tendencies in relation to learner English.
- To describe the characteristics of learner English at different levels of proficiency (e.g. English Profile) and/or at different stages of development (through a longitudinal corpus, e.g. Chen (2013); de Haan and Van der Haagen 2013).

A small sample of research in each of these categories is reviewed below to show the range of research which has been carried out. There is also a short section on findings from the emerging field of learner spoken corpora.

**Learner English (shared L1) and native speaker English**

A comparison of the use of intensifiers by advanced Dutch EFL students and their native-speaking contemporaries was carried out by de Haan and van der Haagen (2013) who found that, compared with their native English speaking contemporaries, their Dutch students initially tended to use in writing intensifiers typical of spoken English such as *really* and *a bit*. Over time, however, these spoken intensifiers disappeared. Reshöft and Gralla (2013) carried out a comparative study of the use of spatial prepositions by German learners of English and native speakers and they
found that native speakers of English used spatial prepositions significantly more often in dynamic contexts than German learners of English. In the field of EAP, Hyland and Milton’s study (1997, cited in Granger 2002) showed that Cantonese learners use a more restricted range of epistemic modifiers than native speakers.

**Learner English (different L1) and native speaker English**

Based on data from the ICLE corpus, Granger (2002) makes a number of observations about general tendencies for learners to overuse, underuse or misuse lexical and grammatical features (in relation to native speakers):

- overuse of adverbs expressing a high degree of certainty, such as *really, of course* or *absolutely*, which are characteristic of speech rather than writing
- underuse of hedging adverbs (e.g. *apparently, possibly, presumably*), which are common in native speaker academic writing (Gilquin, Granger and Paquot 2007: 328)
- overuse of sentence connectors in initial position (e.g. *however*)
- misuse of *on the contrary* as a synonym for *in contrast*
- problems with phraseology, e.g. *as a conclusion* rather than *in conclusion*
- underuse of typical EAP words (*issue, advocate, belief, argument*) and overuse of general and/or vague nouns (*people, thing, problem*).

In general terms, Granger (1994) suggests that many lexical problems are likely to be L1-specific, whereas problems with tenses, articles or prepositions are non-specific to L1.

**Learner spoken corpus findings**

De Cock (2010: 130) cites several examples of learner corpus research including those listed below, which I have selected to illustrate the range of research which can be carried out with learner spoken corpora:

- Discourse markers (e.g. He and Xu 2003).
- Fluency and performance phenomena such as filled and unfilled pauses (e.g. Götz 2007).
- The organization of spoken discourse (e.g. Chen 2004).
- Learners’ underuse of sentential relative clauses in informal contexts, e.g. ‘He’s bought a smartphone, *which is a good idea*, but he can’t use the thing’ (e.g. de Cock 2003).

**Learner English: The developmental sequence**

As we noted earlier, work on the English Profile corpus seeks to describe what learners typically know at the different levels of the CEFR leading to the identification of criterial grammatical features for each level. Hawkins and Buttery
Corpora and the classroom explain that when a grammatical feature is identified as criterial at a given level, it is then regarded as criterial for all subsequent levels, citing two examples of criterial features:

... new verb co-occurrences that appear at B1, such as the ‘ditransitive’ NP-V-NP-NP structure (she asked him his name), are criterial for [B1, B2, C1, C2]; those appearing at B2, for example, the object control structure NP-V-NP-AdjP (he painted the car red), are criterial for [B2, C1, C2].

Similar work in the field of vocabulary is described by Capel (2010: 3) in relation to the wordlists project which aims ‘to establish which words are commonly known by learners around the world at the CEFR levels A1 to B2, and to assign these levels not merely to the words themselves but to their individual meanings’.

While the research carried out for the CEFR is cross-sectional, it will also be useful to look at an example of longitudinal research: Chen (2013) carried out longitudinal research over three years into the progress of Chinese college students with phrasal verbs, using a corpus of argumentative essays. It is interesting that she noted that their use of phrasal verbs in writing actually decreased in the second year.

Applications of learner corpus research

While all the learner corpus research cited above will be of clear interest to second language acquisition researchers, for our purposes we need to highlight the actual and potential pedagogic applications of such research. Probably the most obvious application of corpus data in ELT practice has been in the field of dictionary work. In general terms, Granger (2002: 16–17) argues that the ‘use of megacorpora has made for richer and altogether more useful dictionaries, which provide detailed information on the ranking of meanings, collocations, grammatical patterns, style and frequency’. More specifically, Granger (2002: 19) argues that monolingual learners’ dictionaries can be enhanced by using learner corpus data and cites the Longman Essential Activator Dictionary as the first dictionary to have adopted such an approach. A further example of corpus-informed dictionary work cited by Granger et al. is the integration of the Academic Word List into the Longman Exams Dictionary.

To continue with the theme of vocabulary, Capel (2010) discusses the word lists generated from the Cambridge Learner Corpus (CLC), pointing in particular to the different kinds of searches which can be carried out with the word lists. CLC users can:

- access the words in the list alphabetically
- access words at a specific CEFR level
- identify the CEFR level of given words
- carry out ‘wild card’ searches, e.g. *ing to find all words ending in ‘ing’
- carry out ‘restricted searches on a specific aspect or combination of aspects, including grammar, usage, topic or affixation’.
Vocabulary data from learner corpora can also be put to use in syllabus design. There is perhaps a certain circularity in identifying for inclusion at a given level words which learners typically know at that level, but it can be assumed that these represent a reasonable initial target. Such information can also be put to use in the design of graded readers. Learner corpora may also help us to identify words which learners find difficult and which, therefore, require more attention (Granger 2002). Parallel arguments can be made in relation to the design of a grammar syllabus: if we know which items learners typically find difficult, we can take this into account in sequencing the syllabus and deciding how much time needs to be devoted to specific areas of grammar. The role of learner corpora in syllabus selection is summed up by Meunier (2002: 123, cited in McEnery and Xiao, 2011): ‘... it is important to strike a balance between frequency, difficulty and pedagogical relevance. That is exactly where learner corpus research comes into play to help weigh the importance of each of these’.

Test designers, examiners and materials writers may also benefit from learner corpus data, Hawkins and Buttery (2010: 21) argue, with particular reference to the English Profile Corpus:

[Learner corpus research] can also lead directly to improved publishing outcomes in which not only lexical features are calibrated more closely with textbook preparation at a given level, as in the past, but many grammatical features as well. Textbooks oriented to particular worldwide markets, for example, Chinese or Spanish learners of English, can also benefit from the L1-specific features identified in this research.

The potential for L1-specific CALL programmes is noted by Granger (2002), citing Milton’s WordPilot programme, designed to improve the writing skills of learners in Hong Kong by sensitising them to the typical errors made by Hong Kong learners in their writing. This programme also allows learners to use a reference corpus of native speaker writing for comparison purposes.

In terms of coursebooks, while Burton (2012) notes a general reluctance of coursebook writers to refer to corpus data of any kind, the increasing tendency of coursebooks to align themselves with CEFR levels promises to change this trend: coursebook writers will presumably be guided by the corpus-informed lexical and grammatical specifications of the relevant level.

**Teaching-oriented corpora**

**The need for teaching-oriented corpora**

Thus far much of our discussion has focused on native speaker and learner corpora which have essentially been designed for research purposes. However, as Braun (2005) argues, while these corpora can be exploited for pedagogic purposes, they were not designed for pedagogic purposes. Braun (2005) outlines a number of
limitations of such corpora for classroom purposes, limitations which she then uses as a basis for design criteria for, to use Leech’s (1997) term, ‘teaching-oriented corpora’. These design criteria (Braun 2005) are summarised below:

1. **Size**: in contrast to the multi-million word general corpora, teaching-oriented corpora need only be between 20,000 and 100,000 words so that learners do not drown in data.
2. **‘Intertextual coherence’**: while general corpora deliberately seek a wide range of text types, teaching-oriented corpora should aim to include a more homogenous collection of texts likely to motivate a given set of learners.
3. **Mode**: while it is common for corpora to be available only in text form, it is advantageous for teaching-oriented corpora to be available in audio-visual form as well as text.
4. **Annotation**: general corpora are often tagged for part of speech and/or parsed for the benefit of linguistic research. The annotation of teaching-oriented corpora needs to be pedagogically motivated: Braun (2005: 56), for example, notes that the teaching-oriented corpus, ELISA (English Language Interview Corpus as a Second Language) is annotated with regard to:
   - content-related categories (topics, keywords)
   - L2-related categories (i.e. lexical, grammatical, pragmatic and discourse properties)
   - learner-related categories (level of proficiency, relevant knowledge requirements, skills which can be practised, challenges and difficulties).
5. **Exploitation**: teaching-oriented corpora need to be ‘pedagogically-enriched’, including, for example, guidance notes for teachers, suggested activities and familiarisation tasks for learners. Exploitation of a teaching-oriented corpus should involve both whole-text work as well as corpus analysis work such as concordancing.

**Examples of teaching-oriented corpora**

Brief descriptions of and links to three publicly available teaching-oriented corpora are given below:

1. **ELISA corpus** – English Language Interview Corpus as a Second Language: [http://www.uni-tuebingen.de/elisa/html/elisa_index.html](http://www.uni-tuebingen.de/elisa/html/elisa_index.html)
   The corpus is a collection of video-based interviews with native speakers of different varieties of English (e.g. US, England, Scotland, Ireland and Australia) and from different walks of life who talk about their professional careers. All interviews follow a general pattern, covering a similar range of topics, e.g. what the speakers do, their educational background, how they started their career or business, the type of projects they are involved in, their daily routines and future plans.
   There is an article about this corpus at [http://epubs.surrey.ac.uk/1292/1/fulltext.pdf](http://epubs.surrey.ac.uk/1292/1/fulltext.pdf)
2 SACODEYL Corpus (System-aided compilation and distribution of European Youth Language): http://www.um.es/sacodeyl/
The project includes the collection and distribution of English, French, German, Italian, Lithuanian, Romanian and Spanish teen talk. SACODEYL sees itself as a pedagogical mediator in the language learning process of young Europeans, exploiting web multimedia resources to deliver learning experiences based on data-driven, constructivist approaches to language acquisition.

There is guidance for teachers on how to use the SACODEYL Corpus at http://www.um.es/sacodeyl/data/publications/SACODEYL_guidelines_for_teachers.pdf

3 The BACKBONE corpora: http://projects.ael.uni-tuebingen.de/backbone/moodle/
The BACKBONE suite of pedagogic corpora consists of seven sub-corpora of video-recorded interviews: English, French, German, Polish, Spanish and Turkish as well as French, German, Polish, Spanish and Turkish manifestations of English as a Lingua Franca (ELF).

English is covered by 50 interviews: 25 British and 25 Irish. All the other languages are represented with 25 interviews. The ELF corpus contains a total of 50 interviews, with 10 interviews for each of the base languages French, German, Polish, Spanish and Turkish. The interviews have an average length of 10 minutes. Pedagogic annotation refers to thematic, linguistic and CEFR characteristics; pedagogic enrichment concerns ready-made language learning materials and instructions for exploratory and communicative tasks.

Corpus Search
1 Go to: http://projects.ael.uni-tuebingen.de/backbone/moodle/
2 Click corpora and search.
3 Select the corpora you want to search.
4 Enter a word.
5 Examine the results.

Discussion
How could you make use of this (kind of) corpus in your teaching?
**Improvising corpora for the classroom**

While such teaching-oriented corpora appear to have much to offer, not least as templates for further corpora of this kind, it will probably be beyond the scope of the individual teacher to construct something so elaborate. We need, then, to consider what Tribble (1997) refers to as ‘quick-and-dirty corpora’. As Tribble was writing in 1997, the technical suggestions are not so relevant today: as we noted earlier, with access to the web the construction of a simple corpus is child’s play. However, Tribble’s (1997) rationale for such corpora remains relevant. A fundamental part of Tribble’s rationale for constructing a corpus directly related to learners’ needs is the motivational value of studying relevant texts. He advocates, then, a corpus ‘of expert performances’ (Bazerman 1994: 131) in genres which have relevance to the needs and interests of the learners, noting that this may require the construction of several ‘micro-corpora’. The music corpus discussed in the introduction is an example of this kind of corpus. A slightly different kind of teaching-oriented corpus was suggested by Willis (2003), who used the term ‘pedagogic corpus’. Willis’ (2003) proposal involved compiling the texts learners had used in class into a cumulative corpus which could then be used for grammar and vocabulary work. The advantages of such a corpus, Willis (2003: 166–167) argues, are as follows:

- As the learners have already read or listened to the texts which constitute the corpus, they will be able to understand the context and co-text of specific features chosen for study.
- If the corpus consists of natural texts, as Willis (2003) recommends, then study of a specific feature may well lead to the discovery of other interesting features in the co-text.
- Learners are more likely to recall examples which, having been taken from texts they are familiar with, have real meaning for them.
- The pedagogic corpus approach encourages curiosity and self-reliance.

From a teaching point of view, adopting a pedagogic corpus approach, Willis (2003: 225) emphasises, entails a particular view of second language acquisition:

If we see the pedagogic corpus as central to syllabus and materials design, we can go beyond the view of language learning as the accumulation of a series of language forms. We can see learning as the learner’s growing familiarity with a valuable body of language.

Using learning materials as a corpus is also a feature of Römer’s (2004) text book corpus, a corpus compiled from texts in the (German EFL) coursebooks designed to represent speech e.g. coursebook dialogues. However, Römer (2004) does not describe direct pedagogic applications of this corpus: she discusses this project only in research terms as the aim was to compare textbook representation of certain grammatical features with corpus-attested use of the same features.
A further proposal for a teaching-oriented corpus is made by Allan (2009) who suggests that graded readers can be used to create a corpus: her corpus was made up of Penguin readers at CEF B1 and B2 level, together totalling about 1.5 million words, the majority of which were from contemporary and classic fiction. The main rationale for such an approach is that corpora of authentic texts are too challenging for certain learners:

While more proficient learners may be able to cope with this, those at an intermediate level, situated at B1 or B2 of the Common European Framework of Reference (Council of Europe 2001), for example, are unlikely to be able to deal with the peripheral linguistic content of a search from the BNC or other large corpus.

(Allan 2009: 24)

As a measure of the representativeness of such a corpus, Allan (2009: 30) conducted a comparative study of the frequency of lexical chunks compared with the BNC. While acknowledging that there are some disparities, she concludes that a graded reader corpus can still be of value:

… some chunks in common usage may be screened out in the grading process and due to text genre, occurrences of chunks in the B2 graded corpus may reflect authentic language use quite closely. The size and grading of the B1 corpus, however, does affect access to commonly used chunks, as seen in the case of ‘deal’ and ‘state’.

However, for the teacher looking for a way into DDL with lower-level learners it seems that graded corpora may offer a reasonable balance of accessibility and authenticity in the data it provides.

**Corpus Question**

Build a ‘quick and dirty’ corpus (see Chapter 2 for details) using texts from the internet on a topic of interest to a specific group of learners. When you have your corpus, use antconc (http://www.laurenceanthony.net/antconc_index.html) to carry out some preliminary investigations (e.g. word frequency, ngrams – see Chapter 2 for details).

**Discussion**

How do you think you could exploit your corpus with learners?
Data-Driven Learning (DDL)

**What is data-driven learning?**

DDL is defined by Hadley (2002: 99) thus: ‘Data-driven learning studies vast databases of English text (corpora) with software programs called concordancers, which isolate common patterns in authentic language samples. It is essentially a new form of grammatical consciousness-raising …’ Hadley’s (2002) definition is useful in that it covers the source of data in DDL (corpora) and one kind of learning involved in DDL – grammatical consciousness-raising. We need, however, to be a little more specific about the nature of DDL for our purposes. Data-driven learning offers learners the opportunity to work directly with corpus data, which may come from general corpora such as the BNC or COCA, from specifically designed teaching-oriented corpora, or even from learner corpora. The corpus data for DDL may be provided by the teacher, accessed by the learner under the teacher’s guidance, or accessed by the learner autonomously. In some cases the corpus data may be filtered by the teacher before classroom use, typically by selecting the data which s/he thinks is most suitable for the learners rather than by editing the corpus data itself. In terms of the focus of DDL, while Hadley (2002) refers to grammatical consciousness-raising, DDL arguably works as well – if not better – for lexical or lexico-grammatical information (Allan 2009; Götz 2012). It will be helpful to look at two examples of DDL tasks at this point to illustrate the basic principle of DDL:

**Example 1 (Timmis):**

Look at the concordance lines for *persuade* and *convince* from the academic component of the BNC. Which verb is used more often in the pattern *verb + object + infinitive*?

![BNC concordance: persuade](image)
This is quite a straightforward example of DDL, but it illustrates the fundamental DDL principle of presenting learners with data and asking them to notice linguistic patterns, in this case the typical complementation patterns of two verbs which are closely related semantically. It is important to note here that the corpus data does not tell us that ‘convince someone to do something’ is wrong; it just tells us that it is not frequent. As we observed earlier, corpus data tends to lead to probabilistic rather than deterministic rules.

Example 2:
The DDL task below (Sripicharn 2003) shows how teachers and materials writers can include questions and prompts along with the corpus data to guide learners’ discoveries. In this case, learners use the corpus data to investigate collocations of ‘commit’, with a focus on semantic preference (see Chapter 2).

What do all the underlined words have in common? What do people normally ‘commit’?

How are the actions in line 8 and line 9 different from the others?

1. divorce, yet we know that men commit adultery more often, which suggests
2. their blackness that has made them commit crime, but we cannot ignore the
3. accept the charge that he might commit fraud against Simex by failing to
4. jury to decide if he intended to commit murder and grievous bodily harm.
We should note here that DDL does not require radically different classroom activities: common pedagogic exercise types such as gap-fills, matching activities and multiple choice tasks can be generated from concordance output for DDL purposes.

**What is the rationale for DDL?**

The twin foundations of the rationale for DDL are authenticity and autonomy. One of the main advantages claimed for DDL is that it allows learners to access large amounts of authentic data and, in the form of concordance lines in particular, access to multiple examples of a particular feature (e.g. Chambers 2010; Gilquin and Granger 2010). At first sight, this seems to be in harmony with the greater focus on authentic texts that has often accompanied CLT. This focus has generally been justified on the grounds that authentic materials are preparation for real life and present language in natural contexts, features which make such materials more motivating for the learner. More specifically, Johns (1991) argued that asking learners to study authentic (corpus) data can often lead them to arrive at descriptions of language features which are both more accurate and more practical than those found in ELT materials. The descriptions are more accurate because they are based on real data and more practical because they are expressed in the learner’s own terms. All the kinds of information we have discussed in relation to corpus research are available to the learner, provided they know how to access them, e.g. word and phrase frequency, collocational tendencies, grammar–lexis links.

In terms of autonomy, DDL places responsibility on learners to discover patterns in the data presented. This line of thinking is summed up by Boulton (2009: 37):

> It may even be, in some cases, that learning is more effective without a teacher, i.e. when learners discover things for themselves. This is the basic premise of data-driven learning (DDL), where learners examine naturally-occurring language and discover patterns on their own. DDL is alleged to have many advantages – to foster learner autonomy, increase language awareness, improve ability to deal with authentic language, and so on.
In Johns’ (1991: 2) words, the learner becomes a researcher: ‘... the language-learner is also, essentially, a research worker whose learning needs to be driven by access to linguistic data – hence the term ‘data-driven learning’ (DDL) to describe the approach’.

While it is often the inductive nature of DDL which is stressed, Aston (2001) notes that inductive learning can lead to deductive learning, i.e. hypotheses the learners have derived from the data can then be tested against further data. It is this kind of curiosity-driven corpus search which Aston (2001: 25) considers to be one of the chief potential benefits of DDL: ‘The possibility of encountering and further exploiting such curiosities seems one of the richest potentials of corpora for the language learner, who may acquire a wide variety of linguistic and/or cultural knowledge as a result of such incidental learning’.

The longer-term benefits for learners of such an approach, Aston (2001) argues, can be increased metalinguistic and metacognitive awareness. The potential of DDL for lexical learning is particularly stressed by Allan (2009: 24):

For lexical learning, it is particularly helpful in that it gives learners multiple exposures to words in context, offering potential for deepening word knowledge through the information provided about collocations, contextual behaviour, and register. It would appear to be a valuable explicit ‘focus on form’ technique.

It has also been argued (e.g. Walsh 2010) that corpora can contribute to the development of the four skills. We have already noted that MICASE focuses on academic speaking. In relation to the role of DDL in reading activities, Aston (2001) makes two specific suggestions:

- Concordances of selected words from a text can be used to help learners predict the structure and context of a text and to reinforce key lexis.
- Concordances of the names of characters in a story can be retrieved in order to build a description of the character (Tribble and Jones 1990).

The potential value of corpora for reading is noted by Walsh (2010), who points to the potential use of lexical profiling software in helping teachers and materials writers to assess text difficulty through lexical density measures (i.e. type–token ratios).

Corpora can also be used to improve writing. The British Academic Written English corpus (BAWE), for example, has accompanying materials at http://learnenglish.britishcouncil.org/en/writing-purpose/writing-purpose

Small corpora of student writing, Walsh (2010: 338) observes, can be used to raise learners’ awareness, for example, of overusing certain discourse markers or using a narrow range of vocabulary. The Corpus Search below is an example of a DDL task I have used with non-native PhD students to support academic writing.
Corpus Search

1. Students find an extract from an article they have read and found interesting recently – the abstract, the introduction or the conclusion of the article would be suitable (as these often include the key words for the particular topic).

2. Go to http://www.wordandphrase.info/analyzeText.asp

3. An empty text box will appear.

4. If the article you have chosen is available electronically, copy and paste the abstract into the text box. If it is not available electronically, you will need to type it in.

5. Click Search.

6. Your text will appear in an adjacent box with words colour-coded for frequency.

7. Click the red ‘ACAD’ symbol and frequent academic words will be highlighted.

8. Click on any word that interests you and you will get a list of frequent collocations and a lot of concordance examples.

A simpler kind of word highlighter (AWL) can be found at http://www.nottingham.ac.uk/alzsh3/acvocab/awlhighlighter.htm. This AWL also allows you to create gap-fill exercises around academic words in a given text, as illustrated below:

It would be absurd to suggest that this survey provides a statistically accurate picture of the state of opinion among students and teachers: the sample is but a tiny fraction of the English language learning and teaching population, and questionnaires are not precision instruments. I think, however, that the survey, given the number of responses and relatively wide geographical coverage, can support the following modest conclusions:

<table>
<thead>
<tr>
<th>Gap File produced at level 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>It would be absurd to suggest that this ______ provides a ______ picture of the state of opinion among students and teachers: the sample is but a tiny fraction of the English language learning and teaching population, and questionnaires are not ______ instruments. I think, however, that the ______, given the number of ______ and relatively wide geographical coverage, can support the following modest ______.</td>
</tr>
</tbody>
</table>

The following words will fill the gaps: accurate/consistent/precise/response/statistical/survey/survey

Figure 6.3 Academic word highlighter gap-fill
Among other web-based resources which can help learners with writing are the following:

- Using a programme called Check My Words (http://mws.ust.hk/cmw/index.php), learners can quickly investigate both concordances and collocations of a particular word or phrase.
- Learners can use online text analysis tools (e.g. at http://www.usingenglish.com/resources/text-statistics.php) to examine the lexical density of their texts (i.e. the number of different words relative to the total number of words), the frequency of the words they have used, the average length of sentences and so on.
- For examples of how DDL can be used to help learners make grammatical choices when writing, visit: http://conc.lextutor.ca/concord_writer/index.pl?lingo=English

While it is interesting to consider the way corpora can contribute to skills work, I would argue that, with some exceptions, corpora help more towards building the language resources to facilitate the skill than towards developing the skill itself.

**Reservations about DDL**

Reservations about the learning theory of DDL tend to centre around four areas (Gilquin and Granger 2010): logistical concerns; the teacher’s perspective; the learner’s perspective; the nature of DDL data. Most basically, DDL clearly requires access to corpora, to appropriate software and the ability to use the software. A user-friendly interface is clearly important here. We need also to consider the learning theory behind DDL, which clearly has implications for the roles of teachers and learners. In DDL, teachers are seen as facilitators or guides, helping learners to formulate hypotheses and questions and providing support as they investigate corpus data. This can be a time-consuming process, as Hewings (2012) observes, relating that after a DDL lesson, one learner remarked that he could have told them what they learned from DDL in five minutes. Learners, meanwhile, are seen as active constructors rather than passive recipients of knowledge. In some educational contexts, this change in role may require considerable adaptation on the part of learners and teachers. Furthermore, as we noted in the introduction, KWIC concordances need a different style of reading: reading from the centre out rather than left to right and making sense of a line of text stripped from its original context and co-text. The authentic co-text which contextualises the search item may itself contain vocabulary that is too difficult for learners below advanced level and/or cultural references which are obscure to the learner. We should note here that, while authentic data has been seen as a positive feature of DDL, both the nature and proclaimed benefits of (corpus) authenticity are disputed. While Sinclair (1991: 6) stresses the importance of
Corpora and the classroom 139

studying real language through a botanical analogy, ‘One does not study all of botany by making artificial flowers’. The same analogy was taken up by O’Keeffe (2000) to make a counter-argument:

… in the case of studying real spoken interaction data in the classroom, we are dealing with dried flowers: spoken interactions plucked from their natural contexts and modes (written down etc.). When interactions are ‘dried’ of life, transcribed into print and put before a language learner, there is latent danger of not only misinterpretation semantically, but also pragmatically.

This analogy recalls Widdowson’s (1978: 80) distinction between ‘genuineness’ and ‘authenticity’: ‘Genuineness is a characteristic of the passage itself and is an absolute quality. Authenticity is a characteristic of the relationship between the passage and the reader and it has to do with appropriate response’.

In these terms, corpus data can be seen as ‘genuine’, but not ‘authentic’, especially as, in the form of concordance lines, corpus data appears stripped from context (though the source text with the wider context can sometimes be retrieved).

DDL in perspective

It can be argued, however, that the potential weaknesses outlined above need not be an absolute impediment to DDL. In terms of access to data, we have already made considerable use of the BNC, COCA and MICASE in this volume as they are freely available on the internet. The interfaces are quite user-friendly and the main functions are certainly not beyond the capacity of any teacher or learner to operate. Indeed, Bernardini (2001) urges that the technical challenges of working with corpora should not be overestimated. We have also discussed in this chapter the possibility of developing homemade corpora.

To ensure that concordance lines are accessible, there is no reason, as we have already noted, why the teacher or materials writer should not edit the selection of concordance lines (rather than the concordance line itself) to omit those which might be culturally or linguistically inaccessible to the learners. In terms of contextualising the data, it is often possible to switch from the concordance line to the wider co-text from which the line came with a single keystroke. In the case of spoken data, however, it can be argued that it is both more important and more difficult to reconstruct the context in which an utterance was made.

In relation to the question of whether DDL is accessible to learners at different levels of proficiency, Boulton (2009) argues that careful selection of concordance lines can bring DDL work within reach of lower level learners – while inductive work on concordance lines is cognitively demanding, it is important, as Boulton (2009) emphasises, not to assume that lower level learners are cognitively less efficient than higher level learners. Boulton (2009) also makes the suggestion that DDL work may prove suitable for learners who have struggled with traditional presentation methods.
In assessing the challenge of DDL, we also need to take into account the nature of the data. In this respect, I would argue that Widdowson’s (1978) distinction between ‘genuine’ and ‘authentic’ is actually unnecessary and unhelpful. If we take the position that authentic materials are sometimes helpful and sometimes not, are sometimes better served raw and sometimes edited, then we can work with a simple definition of authenticity: not specifically produced for language teaching purposes. We can also facilitate understanding of the data in other ways: a genre-specific corpus (e.g. a corpus of newspaper texts), as Chambers (2010) notes, is likely to be less challenging than a general corpus as it will have a narrower range of text types. There is also the possibility of using a corpus of ‘comprehensible input’, e.g. a corpus made up of graded readers (Allan 2009) or, indeed, a pedagogic corpus (see ‘Learner corpora’ section earlier in this chapter) to increase the chances of the learner being familiar with the co-text around the search word in the concordance line. Suggestions have also been made that learner corpora could be used in DDL (see also Chapter 7): Meunier (2002), for example, suggests that parallel concordances of the same item in native speaker and learner corpora can help learners in areas of attested difficulty, while Mukherjee and Rohrbach (2006: 225) argue that learners ‘not only profit from the correction of their own mistakes but also from the analysis of their fellow-students’ errors and their corrections’. Meunier (2002) recommends, however, that such work be done judiciously and sparingly and acknowledges that the use of learner data with learners is controversial.

Both Gilquin and Granger (2010) and Chambers (2010) discuss the possibility of a gradual progression with DDL. While Chambers (2010) argues that learners might progress from teacher-prepared concordances to consultation of easy-to-use online resources (e.g. SACODEYL) and then to independent corpus analysis, Gilquin and Granger (2010: 363) note that DDL activities ‘may be located along a cline ranging from teacher-led to learner-led, e.g. from cloze tests to free browsing’. Similarly, Aston (2001) outlines how the learning demands of DDL can be managed by the careful choice of tasks, suggesting that we need to take into account ‘precision’ and ‘recall’ when designing tasks: ‘precision’ refers to the degree to which a search will retrieve only the search item; ‘recall’ refers to the degree to which a search will retrieve all instances of the search item. Entering ‘if’ as a search item to generate conditionals, for example, would have neither high precision nor high recall: it would, for example, generate indirect questions (e.g. I asked him if he knew) as well as the target feature; it would also omit ‘if-less’ conditionals such as, ‘Should you have any questions, let me know’. Aston (2001) suggests a number of other questions we can ask ourselves when assessing the level of challenge of DDL activities:

- Does the task require manipulation of the input by the learners?
- Does the task require all the data to be classified?
- Can the task be done collaboratively?
- Can parts of the task be delegated to stronger learners?
Empirical studies of DDL will clearly be of value in assessing the effectiveness of DDL. Three main kinds of study have been carried out (Gilquin and Granger 2010): attitudinal surveys, research into practices (i.e. how well learners can ‘do’ DDL) and research into learning benefits, though Boulton (2009) points out that research into learning benefits is actually quite scarce. Gilquin and Granger (2010) report that research tends to show mixed results in terms of learners’ attitudes to DDL and ability to ‘do’ DDL while in terms of learning benefits most studies report qualified success. Studies into the kind of DDL tasks which work better are also helpful: Götz (2012), for example, in a small-scale study, found that her learners reacted better to tasks which were very clearly structured by the teacher, i.e. guided discovery tasks rather than fully autonomous discovery tasks. The Corpus Search below, a guided discovery task, is taken from Götz (2012).

### Corpus Search

1. Go to [www.lextutor.ca](http://www.lextutor.ca) and choose *concordance English*.
2. Type in *difficult* as the keyword and include *for* as an optional associated word within four words to the right.
3. Choose BNC Written as your corpus of choice and have a look at the concordance.
4. In which different kinds of patterns is *difficult for* used? Give the abstracted patterns and build categories (e.g. *difficult for sb. (animated NP) to V*ing) and provide an authentic example from the concordance.
5. Now go back and try the same question for a larger database, i.e. choose the fourth option from the drop-down menu, which is ‘all of the above’. With the longer concordance, can you detect more and/or other patterns in which *difficult (for)* can be used? Or do you simply get more confused by too many concordance lines?

In this study, Götz (2012) learners were able to deal with frequency information (e.g. the most common prepositions which follow specified words) more easily than with the task of formulating grammatical rules from the data.

### Final word on DDL

Both the empirical and theoretical arguments for DDL suggest, I would argue, that DDL should be part of the repertoire of teachers and materials writers. Despite these apparent benefits, however, DDL does not seem to have made a widespread impact on language teaching (Braun 2005). This is perhaps a little surprising: nobody has claimed that DDL is a panacea or even that it is suitable for all learners but there is a strong case, I would argue, that, if used *judiciously*, it can be of
significant value to *some* learners. In particular, it is likely to be useful for autonomous learners with very specific needs: PhD students, for example, who may have ‘outgrown’ standard EAP courses (see Chapter 7).

DDL may well, as Aston (2001: 42) outlines, require a shift in mindset on the part of teachers and learners before it enters the mainstream:

… it may be necessary for both teachers and learners to modify many of their presuppositions concerning the nature of language and the nature of language learning. They will need to realize that language use is not simply a matter of employing universally applicable rules; they will need to realize that generalizations from evidence are rarely definitive, but a matter of approximation; and they will need to realize that partial, limited generalizations may be more useful than ones which are broader in scope but harder to apply.

**Discussion**

**What does DDL have to offer in your teaching/learning context?**

Personally, I would argue that the impact of DDL has been limited by an image problem which derives partly from its name: ‘data’ is not a word to inspire most of us, and few people I know in ELT like to feel ‘driven’: Huang’s (2011: 481) term ‘corpus-aided language learning’ seems to fit the bill perfectly, but will it catch on?

**References**


Chapter 7

Corpora and ESP

Introduction

The main aim of this chapter is to provide an introduction to an important aspect of the Corpora/ELT relationship, i.e. the link between corpora and ESP (English for Specific Purposes). Throughout this book, we have discussed the case for a corpus-informed approach to ELT and it can be argued that, in relation to ESP, the case for a corpus-informed approach to teaching and materials design is even stronger, as it is not unusual for English language teachers to find themselves teaching English for specialisms of which they have no particular knowledge. I have myself, for example, taught English to an engineer and to an opera singer, even though I find changing a light bulb and singing a note challenging (especially at the same time). This link between corpora and ESP is likely to remain important for some time to come, given the current patterns of academic and professional mobility which create a need for ESP expertise and materials, so it is an area that clearly deserves our attention from both a research and practitioner perspective.

Three main uses of corpora in ESP are identified by Hewings (2012: 1):

First, in conducting research that underpins effective language teaching and learning. Second, corpora can be used as a way of presenting authentic language data to learners in an approach referred to as ‘data-driven learning’ (DDL). Third, corpora can be a resource for teachers to use in preparing classroom teaching activities and materials.

We can see, then, that this chapter also allows us to revisit, though from a different perspective, themes established earlier in this book: corpus-based language research, data-driven learning, corpus-based materials.

It is relatively straightforward to define ESP. We can, I would suggest, work to the following definition for our purposes: learning English for specific academic, professional or technical purposes. Categorising ESP is, however, a little less straightforward. While it is customary, for example, to talk about EAP (English for Academic Purposes), it is possible to sub-divide this into EGAP (English for General Academic Purposes) and ESAP (English for Specific Academic Purposes)
– you might also want to separate spoken and written academic English. In this chapter, however, as it is an overview, most of our discussion will focus around three broad areas which seem to have attracted the most attention in the ESP corpus field: English for Academic Purposes (EAP), Business English, Engineering English. Hewings (2012) identifies EAP and Business English as two growth areas in the world of ESP corpora, while Engineering English gives us the opportunity to look at a scientific discipline. Our main focus, however, will be on EAP, as this is the area that has enjoyed most attention from researchers, and many of the principles we talk about in relation to EAP will also be applicable to Business English and Engineering English.

**Types of EAP corpus**

Four main types of EAP corpora are identified by Nesi (2014). While she is in this instance speaking specifically about EAP corpora, I would argue that the categories she uses seem to apply equally well to other kinds of ESP corpus. We will look in more detail at each type in this chapter, but they are briefly described below:

- corpora of ‘expert’ writing
- learner corpora
- corpora of university student writing
- spoken academic corpora.

Corpora of expert writing are, Nesi (2014) points out, the most common type because expert writing is available in the public domain and, therefore, easy to access for the purposes of corpus compilation. While it is more difficult to obtain student writing than expert writing for corpora, student writing is important as it may present a more realistic and attainable target for learners (Nesi 2014) – it is also useful for the purposes of comparison with expert corpora. We could add to this list ‘pedagogic corpora’ in the sense that we defined it in the previous chapter: Ward’s engineering corpus (see below), for example, was constructed from engineering textbook materials.

The idea of ‘quick and dirty’ or, to use the more charitable term, ‘ad hoc corpora’ we discussed in the previous chapter is also highly relevant here: there are many areas of ESP, not catered for by published corpora, where it will be expedient to construct ad hoc corpora for our own classes. By way of example, a PhD student of mine is about to compile a corpus of football English to help foreign footballers who come to play for English clubs. It is also worth adding to the list of available types corpora constructed autonomously by learners in areas of specific interest to them, as proposed by Lee and Swales (2006) and Gilquin, Granger and Paquot (2007). If we add these corpora types to Nesi’s (2014) categories (above), we can arrive at a classification that, I believe, covers the possibilities for ESP:
1 Written ESP corpora (expert or student writing).
2 Spoken ESP corpora (expert or student speech).
3 Learner academic corpora (spoken or written).
4 Pedagogic corpora (e.g. corpora of textbook materials).
5 Ad hoc corpora (i.e. quickly produced for a specific context).
6 Autonomously constructed corpora (i.e. learners play a role in the construction of the corpus).
7 Sub-corpora of large corpora (e.g. the academic component of the British National Corpus).

This chapter deals with each of the three ESP domains (EAP, Business English, Engineering English), though, for the reasons noted above, our emphasis is on EAP. We also consider briefly learner EAP corpora. Finally, we look at the implications of ESP corpus research in terms of syllabus, methodology and materials.

**Examples of EAP corpora**

Below is a brief description of four EAP corpora, two written and two spoken corpora, to give some idea of the range of EAP corpora which have been produced. Each corpus is described in terms of size and content (e.g. the nature of the data, the contributors or the disciplinary areas covered). There is also information on how to access the corpus (if it is possible) and where to get further information.

1 **British Academic Written English Corpus (BAWE)**
   Size: 6,506,995 words
   Content:
   - 3,000 good-standard student assignments.
   - Mostly (but not exclusively) native speakers with a grade equivalent to an upper second or first class honours degree.
   - Disciplines: four broad disciplinary areas (Arts and Humanities, Social Sciences, Life Sciences and Physical Sciences).
   - Four levels of study (undergraduate and taught masters level).

   Availability: Free access
   Further information: [http://www.coventry.ac.uk/research/research-directory/art-design/british-academic-written-english-corpus-bawe/](http://www.coventry.ac.uk/research/research-directory/art-design/british-academic-written-english-corpus-bawe/)
2 MICUSP (Michigan Corpus of Upper Level Student Papers)
Size: 1.6 million words
Content:

- A-grade papers or ungraded papers that have been assessed and accepted (such as research proposals), but not published (the length of the texts ranges from 500 to 10,000 words).
- Writers at different stages of undergraduate and graduate level study, both native and non-native speakers, from across the University [of Michigan].
- Disciplines: Humanities, Social Sciences and Education, Physical Sciences and Engineering, Biological and Health Sciences.

Availability: Subscription only
Further information: http://legacyweb.lsa.umich.edu/eli/eli1/micusp/Index.htm

3 BASE (British Academic Spoken English)
Size: 1,644,942 words
Content:

- 160 lectures and 40 seminars recorded in a variety of departments.
- Lecturers and students are included.
- Disciplines: Arts and Humanities, Life Sciences, Physical Sciences, Social Sciences.

Availability: Texts freely available
Further information: http://www2.warwick.ac.uk/fac/soc/al/research/collect/base/

4 ELFA (English as a Lingua Franca in Academic Settings)
Size: 0.9 million words of transcribed speech (recorded at the University of Tampere and Tampere University of Technology, Finland)
Content:

- Lectures, seminars, thesis defences, conference presentations and other academic speech events.
- Non-native speaker lecturers and students.

Availability: Freely available (with licence)
Further information: http://www.helsinki.fi/englanti/elfa/project.html
Corpus Search

1. Go to: https://the.sketchengine.co.uk/open/
2. Select BASE corpus.
3. Enter *point* in the search box.
4. Click KWIC on the side bar.
5. Examine the results.

Discussion

Which of the collocations and/or lexical chunks of *point* identified in your search would you highlight for EAP students? What different meanings of *point* can you identify?

Corpus Search

Repeat steps 1 to 5 above, but select BAWE at step 2 and choose your own search word at step 3.

From this brief overview, we can see some of the main design issues for EAP corpora. The most obvious is size: specialised corpora are normally smaller than general corpora (although the Pearson International Corpus of Academic English comprises 36 million words). In addition to size, the overall organisation of the corpus is clearly important: we need to decide whether to organise the corpus by discipline or by genre or by speech event or some combination of these factors. The choice of contributors is also a central issue as this will determine the purposes for which the corpus can be used. Contributors might, for example be established academic experts, students of a particular discipline or EAP learners. And in each of these categories, we could make a native/non-native distinction.

- Experts in the field
- Students/Novices in the field
- Native speakers
- Non-native speakers (expert users)
- Non-native speakers (learners).
English for Academic Purposes corpus research

The contribution of corpora to EAP research is summarised by Gilquin, Granger and Paquot (2007: 320) thus: ‘[corpus-linguistic] studies make two significant contributions to the field of EAP: detailed descriptions of its distinctive linguistic features, and more specifically its highly specific phraseology, and careful analyses of linguistic variability across academic genres and disciplines’.

English for Academic Purposes corpora and lexis

We will focus on headline findings in EAP research in this section, beginning with lexical frequency in academic English, which has enjoyed considerable attention in recent years (e.g. Coxhead 2011; Ackermann and Chen 2013; Simpson-Vlach and Ellis 2010; Gardner and Davies 2014). Such frequency lists, Gardner and Davies (2014: 306) argue, can serve a number of purposes:

- establishing vocabulary learning goals
- assessing vocabulary knowledge and growth
- analysing text difficulty and richness
- creating and modifying reading materials
- designing vocabulary learning tools
- determining the vocabulary components of academic curricula.

EAP corpora and word frequency

The pioneering project in the field was the Academic Word List (Coxhead 2000), developed from the 3.5 million word Academic Corpus. The data for the Academic Corpus was drawn from four broad disciplinary areas: Arts, Commerce, Law and Science. The source documents were journal articles, book chapters, course workbooks, laboratory manuals, and course notes, i.e. it is a written corpus. The Academic Word List is based on frequency and consists of 570 word families (e.g. analyse, analysis and analyst would constitute one word family). It excludes words, however, which are among the 2,000 most common words in general English (Coxhead 2011). The exact criteria for inclusion in the AWL are explained by Coxhead (2011: 356): ‘The word families had to occur 100 times or more in each of the four disciplines of the corpus (frequency), in 15 or more of the subject areas (range), and over 10 times in the four disciplines (uniformity)’.

A more recent academic word list has been produced and described by Gardner and Davies (2014). Their Academic Vocabulary List is based on the 120 million word academic component of COCA. The AVL reflects nine academic disciplines, with 85 million words taken from academic journals. The corpus also includes, however, around 31 million words from lighter academic magazines and 7.5 million words from the business sections of newspapers. While acknowledging the seminal role of Coxhead’s (2000) Academic Word List, Gardner and Davies
Corpora and ESP (2014) point out a number of features which distinguish their list: The COCA corpus from which their AVL list is derived is grammatically tagged; the AVL deals with lemmas rather than word families; the AVL does not refer to what Gardner and Davies (2014) consider to be the rather outdated General Service List. Gardner and Davies (2014: 312) summarise the main criteria which governed the design of the AVL:

1. The new list must initially be determined by using lemmas, not word families. Subsequent groupings of the list into families may be warranted for certain instructional and research purposes.
2. The new list must be based on a large and representative corpus of academic English, covering many important academic disciplines.
3. The new list must be statistically derived (using both frequency and dispersion statistics) from a large and balanced corpus consisting of both academic and non-academic materials. The corpus must be large enough and the statistics powerful enough to be able to separate academic core words (those that appear in the vast majority of the various academic disciplines) from general high-frequency words (those that appear with roughly equal and high frequency across all major registers of the larger corpus, including the academic register), as well as from academic technical words (those that appear in a narrow range of academic disciplines).
4. The academic materials in the larger corpus, as well as the non-academic materials to which it will be compared, must represent contemporary English, not dated materials from 20 to 100 years ago. Otherwise, the validity of the new list could be questioned.
5. The new list must be tested against both academic and non-academic corpora, or corpus-derived lists, to determine its validity and reliability as a list of core academic words.

It is the application of these criteria which leads Gardner and Davies (2014: 325) to claim that the AVL is ‘the most current, accurate, and comprehensive list of core academic vocabulary in existence today’.

The usefulness of a general academic word list (which applies to both the AWL and the AVL) has, however, been questioned by Hyland and Tse (2009) who argue that much academic language is discipline-specific even to the extent that the same word can have slightly different meanings or uses in different disciplines (‘visualisation’ as used by sports psychologists, for example, has a different meaning from how it is used by Tomlinson when discussing reading techniques). One potential solution to the problem of building learners’ discipline-specific vocabulary, Coxhead (2011) suggests, is to build discipline-specific frequency lists on top of general English frequency lists and general academic English frequency lists.
The Table 7.1 presents the 60 most frequent word families in the Academic Word List (http://www.victoria.ac.nz/lals/resources/academicwordlist/most-frequent/sublist-1-most-frequent-words). The most frequent members of the word families in Sublist 1 are listed below.

Table 7.1 Academic word list (sublist 1) (Coxhead 2000)

| analysis | established | occur |
| approach | estimate | percent |
| area | evidence | period |
| assessment | export | policy |
| assume | factors | principle |
| authority | financial | procedure |
| available | formula | process |
| benefit | function | required |
| concept | identified | research |
| consistent | income | response |
| constitutional | indicate | role |
| context | individual | section |
| contract | interpretation | sector |
| create | involved | significant |
| data | issues | similar |
| definition | labour | source |
| derived | legal | specific |
| distribution | legislation | structure |
| economic | major | theory |
| environment | method | variables |

Discussion

How do you think this list could be exploited pedagogically?

While acknowledging the potential usefulness of such frequency lists, we need to keep in mind the learner perspective, i.e. evidence of which words that specific groups of learners find difficult is an important element in deciding what to teach and in what order. We also need to look beyond the frequency of individual words to see what categories of word can be particularly useful: Aktas and Cortes (2008), for example, focus on the usefulness of shell nouns in organising discourse, e.g. words such as problem, issue, debate which can refer to a stretch of discourse in the text. In the short extract, below (from this book), for example, issue relates to all the previous discussion of the spoken language syllabus:

Aktas and Cortes (2008), cited in Coxhead (2010) found that, while L2 writers used shell nouns, they did not use them for the same cohesive purposes as L1 writers.
We discussed in general terms the importance of collocation for language production in Chapter 2, and the importance of a stock of collocations for academic writing is underlined by Haswell (1991: 236, cited in Hyland 2008: 5), who notes that experienced academic writers rely on collocations significantly more than apprentice writers. Work on academic collocations is, then, of clear interest to us. The compilation of an academic collocation list (ACL) is described in some detail by Ackermann and Chen (2013), who produced such a list from the 37 million word PICAE (Pearson International Corpus of Academic English). Ackermann et al. (2011) provide the following overview of PICAE: ‘The corpus includes curricular English as found in lectures, seminars, textbooks and journal papers. It also samples extracurricular English that students will encounter, from university administration to transcripts of radio broadcasts. It is mainly sourced from the web’.

The ACL comprises only lexical collocations, i.e. collocations consisting of two lexical words (e.g. strong argument) rather than a lexical word and a grammatical word (e.g. depend on). Ackermann and Chen (2013) contrast this focus on lexical collocation with Durrant’s (2009) collocation frequency list, which also included grammatical collocations, pointing out that the top five collocations in Durrant’s (2009) list are all grammatical collocations: this study, associated with, based on, and respectively, due to. While Ackermann and Chen (2013: 246) acknowledge that these collocations may be of some value to learners, they explain their emphasis on lexical collocations thus:

[lexical collocations] contain certain variability and are thus more dynamic, while grammatical collocations or idioms consist of comparatively fixed patterns and are consequently more predictable. The former is more challenging for learners to master whereas the latter [grammatical collocations] can generally be treated as holistic units and learners can more easily internalize the usage into their lexicon.

In their study, Ackermann and Chen (2013: 238) focused on the following categories of lexical collocations: verb/noun (e.g. gather data), adjective/noun (e.g. systematic approach), adverb/adjective (e.g. increasingly complex), and adverb/verb (e.g. significantly affect). The selection process they adopted for the selection of collocation for the ACL went beyond automatic statistical identification of collocations and comprised four main stages (Ackermann and Chen 2013: 235):

1 computational analysis
2 refinement of the data-driven list based on quantitative and qualitative parameters
Expert review and systematization.

The human intervention at stages two, three and four, Ackermann and Chen (2013) argue, is crucial to ensuring that the list is of maximum pedagogical value. Indeed, Ackermann and Chen (2013) stress the pedagogic orientation of their research:

> Focusing on lexical collocations only, we present a new Academic Collocation List compiled using a mixed-method approach of corpus statistics and expert judgement, consisting of the 2,468 most frequent and pedagogically relevant entries we believe can be immediately operationalized by EAP teachers and students.

The list can be accessed at http://pearsonpte.com/research/academic-collocation-list/

**Discussion**

**How could this list be organised for pedagogic purposes?**

Having covered academic word and collocation frequency, it is a natural step to consider the frequency of lexical chunks in an academic context: we established in general terms the importance of lexical chunks for fluent production in Chapter 2, particularly in relation to the work of Martinez and Schmitt (2012). We consider now academic lexical chunks, or, to use the terminology of Simpson-Vlach and Ellis (2010) ‘academic formulas’. For their research into academic formulas, Simpson-Vlach and Ellis (2010) used the following corpora:

- **Academic speech:** MICASE (1.7 million words); BNC spoken academic component (431,000 words).
- **Academic writing:** Hyland’s (2004) research article corpus (1.2 million words); BNC written academic files (931,000 words).
- **Comparison Corpora:** for non-academic speech, the Switchboard Corpus (2.9 million words); for non-academic writing, the FLOB and Frown corpora (1.9 million words).

The selection criteria for the inclusion of items on the academic formulas list (AFL) are set out by Simpson-Vlach and Ellis (2010: 487) thus: ‘The AFL includes formulaic sequences identified as: (i) frequent recurrent patterns in corpora of written and spoken language; which (ii) occur significantly more often in academic than in non-academic discourse; and (iii) inhabit a wide range
of academic genres’. The selection criteria are not, however, purely statistical. As was the case with the Academic Collocations List (above), human judgement plays an important part in refining the list. This judgement is integrated into a calculation of ‘formula teaching worth’ to indicate teaching priority. It is this focus on ‘teaching worth’ that in part distinguishes the approach from the lexical bundle approach adopted by Biber et al. (1998, 2004), as Simpson-Vlach and Ellis (2010: 490) explain:

The lexical bundle approach of Biber and colleagues (1998, 2004), based solely on frequency, has the advantage of being methodologically straightforward, but results in long lists of recurrent word sequences that collapse distinctions that intuition would deem relevant. For example, few would argue with the intuitive claim that sequences such as ‘on the other hand’ and ‘at the same time’ are more psycholinguistically salient than sequences such as ‘to do with the’, or ‘I think it was’, even though their frequency profiles may put them on equivalent lists.

The academic formulas are also categorised into three broad areas of pragmatic function to facilitate inclusion in an EAP syllabus: referential expressions, stance formulas and discourse-organising functions. The categorisation is quite complex, however, as each of these three broad areas is divided into subcategories. The outline categorisation of the AFL (Simpson-Vlach and Ellis 2010) is given below.

- **Group A: Referential expressions** The largest of the three major functional groupings, the referential expressions category encompasses five subcategories: specification of attributes, identification and focus, contrast and comparison, deictics and locatives, and vagueness markers.
- **Group B: Stance formulas** Stance formulas include six functional subcategories: hedges and evaluative formulas, epistemic stance, obligation and directives, ability and possibility, intention/volition.
- **Group C: Discourse-organizing expressions** Discourse organizers in the AFL fall into four main subcategories: metadiscourse, topic introduction, topic elaboration and discourse markers. Each of these functions involves either signalling or referring to prior or upcoming discourse.

The most frequent phrases from this list are shown below (and available at http://www-personal.umich.edu/~ncellis/NickEllis/Publications_files/AFL_paper_AppLinxPrepub.pdf).

1. on the other hand
2. due to the fact that
3. on the other hand the
4. it should be noted
it is not possible to
a wide range of
there are a number of
in such a way that
take into account the
as can be seen
it is clear that
take into account
can be used to
in this paper we
are likely to
in the next section
a large number of
the United Kingdom
on the basis of the
that there is no
over a period of
as a result of the

Discussion

How could this phrase frequency list be exploited pedagogically? Which phrases seem to you to be particularly useful (or not so useful)?

We progress now from lexical chunks as academic formulas to lexical bundles, recalling (Chapter 3) that lexical bundles are commonly recurring strings of a specified length (e.g. 3-word, 4-word or 5-word sequences) which may or may not be meaningful units, e.g. in the end is a meaningful 3-word bundle (and so also a lexical chunk in our terms); some of the is not a meaningful bundle in itself. Hyland (2008: 4) carried out research into the use of bundles in academic texts using ‘a 3.5 million word corpus of research articles, doctoral dissertations and Master’s theses in four disciplines to learn something of disciplinary variations in their frequencies and preferred uses’. The table below (Hyland 2008: 7) shows the most frequent 3-, 4- and 5-word bundles in his corpus.
Table 7.2 Academic bundles frequency list (adapted from Hyland 2008)

<table>
<thead>
<tr>
<th>3-word</th>
<th>4-word</th>
<th>5-word</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 in order to</td>
<td>1629 on the other hand</td>
<td>726 on the other hand the</td>
</tr>
<tr>
<td>2 in terms of</td>
<td>1203 at the same time</td>
<td>337 at the end of the</td>
</tr>
<tr>
<td>3 one of the</td>
<td>1092 in the case of</td>
<td>334 it should be noted that</td>
</tr>
<tr>
<td>4 the use of</td>
<td>1081 the end of the</td>
<td>258 it can be seen that</td>
</tr>
<tr>
<td>5 as well as</td>
<td>1044 as well as the</td>
<td>253 due to the fact that</td>
</tr>
<tr>
<td>6 the number of</td>
<td>992 at the end of</td>
<td>252 at the beginning of the</td>
</tr>
<tr>
<td>7 due to the</td>
<td>886 in terms of the</td>
<td>251 may be due to the</td>
</tr>
<tr>
<td>8 on the other</td>
<td>810 on the basis of</td>
<td>247 it was found that the</td>
</tr>
<tr>
<td>9 based on the</td>
<td>801 in the present study</td>
<td>225 to the fact that the</td>
</tr>
<tr>
<td>10 the other hand</td>
<td>730 is one of the</td>
<td>209 there are a number of</td>
</tr>
</tbody>
</table>

Discussion

How could this list be exploited pedagogically?

While Table 7.2 reflects overall frequency of the lexical bundles in his corpus, Hyland (2008) found that frequency varied according to academic discipline, which recalls our discussion earlier in this chapter about discipline-specific frequency at the level of academic words. Hyland (2008: 122–123) drew up a functional categorisation of the 4-word bundles he had identified, suggesting three main categories:

- Research-oriented bundles which ‘help writers to structure their activities and experiences of the real world’.
- Text-oriented bundles which are ‘concerned with the organisation of the text and its meaning as a message or argument’.
- Participant-oriented bundles which are ‘focused on the writer or reader of the text’.

Each of these categories, however, has sub-divisions: Hyland’s (2008: 122–123) categorisation is reproduced below.

**Research-oriented – help writers to structure their activities and experiences of the real world**

- **Location**: indicating time/place (*at the start of, at the same time, in the present study*).
• **Procedure**: *(the use of the, the role of the, the purpose of the, the operation of the).*
• **Quantification**: *(the magnitude of the, a wide range of, one of the most).*
• **Description**: *(the structure of the, the size of the, the surface of the).*
• **Topic**: related to the field of research *(in the Hong Kong, the currency board system).*

**Text-oriented** – concerned with the organisation of the text and its meaning as a message or argument

• **Transition signals**: establishing additive or contrastive links between elements *(on the other hand, in addition to the, in contrast to the).*
• **Resultative signals**: mark inferential or causative relations between elements *(as a result of, it was found that, these results suggest that).*
• **Structuring signals**: text-reflexive markers which organise stretches of discourse or direct reader elsewhere in text *(in the present study, in the next section, as shown in fig.).*
• **Framing signals**: situate arguments by specifying limiting conditions *(in the case of, with respect to the, on the basis of, in the presence of, with the exception of).*

**Participant-oriented** – focused on the writer or reader of the text

• **Stance features**: convey the writer’s attitudes and evaluations *(are likely to be, may be due to, it is possible that).*
• **Engagement features**: address readers directly *(it should be noted, as can be seen).*

### Discussion

*(How) do you think such a categorisation of academic lexical bundles could help you and your learners?*

**English for Academic Purposes corpora and genre analysis**

While corpus-based genre research has not been carried out on quite the same scale as lexico-grammatical research in EAP, it will be useful to consider two examples of research at the level of genre. Lau (2004, cited in Hsieh and Liou 2008) compared research abstracts written by 50 EFL PhD students with research abstracts written by 30 international scholars in life sciences journals. The abstracts were analysed in relation to a prototypical move structure for abstracts of: Introduction–Purpose–Methods–Results–Conclusion. The research abstracts
produced by the scholars, Lau (2004) points out, were generally closer to the prototypical sequence than those of the students. Lau (2004, cited in Hsieh and Liou 2008) found three common types of move structure in the students’ abstracts which deviated from the prototype:

a Background–Purpose–Results–Conclusion
b Background–Results–Conclusion
c Results–Conclusion.

Hsieh and Liou (2008) compared the move structure of abstracts written for applied linguistics journals with those written for local applied linguistics conferences. The abstracts were assessed in relation to a prototypical move sequence of Background–Gap–Purpose–Method–Results–Conclusion.

Four different types of deviation from the prototype were identified by Hsieh and Liou (2008):

1 Inverted move sequence, i.e. a change in the prototypical six-move sequence of Background–Gap–Purpose–Method–Results–Conclusion.
2 Missing obligatory moves, i.e. the omission of one of the moves from the sequence.
3 Disproportional abstracts, i.e. when too much space is devoted to one move in the sequence.
4 Outlining information in a move, i.e. the use of numbers or letters to list information in a move.

Hsieh and Liou (2008: 37) conclude that ‘the writers of the conference paper abstracts were not as familiar with Western abstract-writing conventions as the writers of the journal article abstracts’. For our pedagogically-oriented purposes, it is also interesting that Hsieh and Liou (2008: 33–34) linked genre analysis with lexico-grammatical analysis to show how genre moves were typically lexicalised in the Journal Paper Corpus:

Table 7.3 Common phraseology used in the moves in the JPA Corpus (adapted from Hsieh and Liou 2008)

<table>
<thead>
<tr>
<th>Move</th>
<th>Examples</th>
</tr>
</thead>
</table>
| Background | • This article presents the results of …  
<p>|            | • several studies have shown …                                               |
| Gap        | • But, to date, very few …                                                 |
|            | • … however …                                                              |
| Purpose    | • The goal/aim of the study is …                                           |
|            | • this article presents evidence of/reports on the findings of …            |
|            | • the survey concerned …                                                    |
|            | • this study investigated/examined/comparisons/presents/explored …         |</p>
<table>
<thead>
<tr>
<th>Move</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results</td>
<td>• The results showed/revealed/suggested that …</td>
</tr>
<tr>
<td></td>
<td>• findings of the study indicated …</td>
</tr>
<tr>
<td></td>
<td>• … were found to …</td>
</tr>
<tr>
<td>Method</td>
<td>• The analysis is based on …</td>
</tr>
<tr>
<td>Conclusion</td>
<td>• The research results provide …</td>
</tr>
<tr>
<td></td>
<td>• it is argued that …</td>
</tr>
<tr>
<td></td>
<td>• these findings lend support to …</td>
</tr>
<tr>
<td></td>
<td>• … findings demonstrate/imply …</td>
</tr>
<tr>
<td></td>
<td>• this study reveals that …</td>
</tr>
<tr>
<td></td>
<td>• it is proposed/posited that …</td>
</tr>
<tr>
<td></td>
<td>• the results may be/are interpreted</td>
</tr>
</tbody>
</table>

**Discussion**

(How) do you think this list could be exploited pedagogically?

The corpus of English as a Lingua Franca in Academic settings has produced a significant amount of research on various aspects of EAP. The motivation for such research is the argument that English has become established as an academic lingua franca for publications and is increasingly the medium of instruction in universities outside the ‘inner circle’ of native-speaking countries. Jenkins, Cogo and Dewey (2011: 300) cite the following example of ELF research in an academic context that looks at the use of progressive forms in an ELF context:

Ranta (2006) draws on the ELFA corpus in her analysis of the ways in which ELF speakers extend their use of the progressive in linguistic contexts where native English would require the simple form. One of these is stative verbs. Ranta’s (2006) data show ELF speakers using progressive forms such as are belonging rather than the corresponding simple form, belong; another context is that of habitual activity, where ELF speakers refer, for example, to the air that we are breathing, rather than the air that we breathe.

This use of the progressive, Jenkins, Cogo and Dewey (2011) remark is strategic in that it makes the statement more prominent and thus gains the attention of the interlocutor.

**Engineering corpora**

**Pedagogic engineering corpora**

We begin our discussion of engineering corpora with an example of developing a pedagogic engineering corpus in relation to a real-world problem (Ward 2009): it is worth this example of a pedagogic corpus in some detail, if we keep in mind that the principles could be applied to the development of corpora for other disciplines
in other contexts. Indeed, it seems to me that this case study is a model of how a corpus can be produced that is somewhat more elaborate than an ad hoc corpus, but less time-consuming and expensive to produce than large scale projects such as the Hong Kong Engineering Corpus. It also demonstrates the principle of fitness for purpose in relation to corpus design and exploitation.

Ward’s (2009) point of departure for his corpus project was the observation that many engineering graduates in developing nations begin their university courses with a vocabulary far short of the level required to read academic material in English. These students, Ward (2009) argues, are not in a position to follow a rather idealised sequence of learning first the 2,000 most common words in English, then the 570 word families on the Academic Word List (Coxhead 2000) before progressing to the discipline-specific vocabulary of engineering. The challenge Ward (2009: 171) set himself was to ‘create a word list as a basis for a lexical syllabus’ for pre-university engineering students. The word list had to meet two fundamental criteria (Ward 2009: 171):

1. It had to be useful enough in terms of text coverage and general frequency, for engineers in all sub-disciplines of engineering.
2. It had to be easy enough to work with, in terms of length and technicality, for learners who have nothing like mastery of the GSL (General Service List) or the AWL (Academic Word List).

The stages in the process of developing the corpus are summarised below, based on Ward (2009):

1. Lecturers in five engineering faculties (chemical, civil, electrical, industrial and mechanical) were contacted and asked each to supply the names of five textbooks commonly used with third and fourth year undergraduate studies.
2. Pages were selected at random from each textbook until a total of 10,000 words had been collected from each book. This produced a corpus of over 250,000 words (i.e. 25 text books × 10,000 word sample).

The resulting corpus was then used to generate an engineering word list. Three criteria were applied in drawing up the list:

- Function words were excluded.
- The basic item was a word type (i.e. not a lemma or word family).
- Words were only included if they occurred five times or more in each subsection (i.e. chemical, civil, electrical, industrial and mechanical).

An engineering list of 299 words was thus generated (many of which also occur in the GSL or AWL). The ten most frequent words are listed below to illustrate Ward’s (2009) that the words are mostly sub-technical rather than highly specialised.

1. system
2. shown
We will consider one further example of a pedagogic engineering corpus, again with a focus on the process of construction. Ng et al. (2013) compiled a pedagogic engineering corpus from Malaysian fourth and fifth year engineering technology textbooks. They then produced an Engineering Technology Word List by applying the following criteria (Ng et al. 2013: 48):

1. The frequency of the words from the engineering technology corpus must be significant enough, according to the Keyword function in the WordSmith Tools 5.0 to be regarded as the specialised engineering vocabulary after a comparative analysis with a larger reference corpus (BNC).
2. The words must be closely related to the engineering field. Thus, the lexical items found must be outside of the GSL and AWL list of words. The general English function words must also be removed.
3. The specialised words identified must match the technical words found in the McGraw-Hill online [technical] dictionary to be considered as technical words. Else, the particular word should be regarded just as a semi-technical engineering word.
4. The identified technical and semi-technical words must be cross-checked with the two experts in the field of engineering to ensure the appropriate words were regarded as the technical words or the semi-technical words or should be omitted.

There are similarities here with Ward’s (2009) – the discipline; the use of textbooks and the omission of function words – but there are also interesting refinements to note:

- The use of a reference corpus to establish ‘keyness’.
- The exclusion of words from the GSL and AWL.
- Consultation of a technical dictionary to validate the selection.
- Consultation with experts to validate the selection.

We have moved a step nearer a fully-fledged publicly-available engineering corpus, to which we now turn our attention.
The Hong Kong Engineering Corpus

As an example of a large ESP corpus, which is free to access, we can take the Hong Kong Engineering Corpus (rcpce.engl.polyu.edu.hk/HKEC). Cheng (2010: 70) provides the following description of this corpus:

The Hong Kong Engineering Corpus … is comprised of 1,066,602 words and is the product of the first large-scale research project to collect corpus texts representative of the English language of the engineering sector in Hong Kong … The Hong Kong Engineering Corpus consists primarily of the texts retrieved from the following sources: the Hong Kong Institution of Engineers (HKIE) website, the CD-ROM holding the 10th Anniversary HKIE Transactions (a professional journal), newsletters from the i-version Journal Hong Kong Engineer Online, and other engineering-related websites, primarily Hong Kong government departments, academic institutions, and engineering companies.

Corpus Search

Carry out the search below at rcpce.engl.polyu.edu.hk/HKEC, but before you do, try to guess what the most frequent right-hand collocate of energy will be:

A discussion of HKEC research is a good opportunity to remind ourselves of concgrams (see Chapter 3) as this corpus has a ‘concgram’ function. Warren (2010: 169) describes this function of HKEC:
The other significant innovation [with HKEC] is the software, concgramOnline, which allows users to go beyond concordances of individual words and clusters. It does this by automatically identifying instances of phraseological variation when two or three words or phrases co-occur, irrespective of either constituency or positional variation.

The most frequent concgrams from HKEC are reproduced below:

<table>
<thead>
<tr>
<th>RANK</th>
<th>WORD</th>
<th>CO-OCCURRING WORD</th>
<th>CO-OCCURRING WORD INSTANCES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hong</td>
<td>Kong</td>
<td>37928</td>
</tr>
<tr>
<td>2</td>
<td>as</td>
<td>such</td>
<td>8170</td>
</tr>
<tr>
<td>3</td>
<td>as</td>
<td>well</td>
<td>6241</td>
</tr>
<tr>
<td>4</td>
<td>less</td>
<td>than</td>
<td>4705</td>
</tr>
<tr>
<td>5</td>
<td>energy</td>
<td>efficiency</td>
<td>4559</td>
</tr>
<tr>
<td>6</td>
<td>more</td>
<td>than</td>
<td>4357</td>
</tr>
<tr>
<td>7</td>
<td>carried</td>
<td>out</td>
<td>3854</td>
</tr>
<tr>
<td>8</td>
<td>air</td>
<td>quality</td>
<td>3559</td>
</tr>
<tr>
<td>9</td>
<td>as</td>
<td>should</td>
<td>3234</td>
</tr>
<tr>
<td>10</td>
<td>million</td>
<td>HK</td>
<td>3128</td>
</tr>
<tr>
<td>11</td>
<td>environ</td>
<td>protection</td>
<td>2742</td>
</tr>
<tr>
<td>12</td>
<td>as</td>
<td>department</td>
<td>2680</td>
</tr>
<tr>
<td>13</td>
<td>as</td>
<td>shall</td>
<td>2641</td>
</tr>
<tr>
<td>14</td>
<td>quality</td>
<td>Water</td>
<td>2417</td>
</tr>
<tr>
<td>15</td>
<td>as</td>
<td>consumption</td>
<td>2291</td>
</tr>
<tr>
<td>16</td>
<td>as</td>
<td>CO</td>
<td>2278</td>
</tr>
<tr>
<td>17</td>
<td>as</td>
<td>code</td>
<td>2267</td>
</tr>
<tr>
<td>18</td>
<td>as</td>
<td>may</td>
<td>2224</td>
</tr>
<tr>
<td>19</td>
<td>supply</td>
<td>water</td>
<td>2171</td>
</tr>
<tr>
<td>20</td>
<td>University</td>
<td>Hong</td>
<td>2031</td>
</tr>
<tr>
<td>21</td>
<td>electrical</td>
<td>mechanical</td>
<td>2045</td>
</tr>
<tr>
<td>22</td>
<td>as</td>
<td>possible</td>
<td>1982</td>
</tr>
<tr>
<td>23</td>
<td>as</td>
<td>quality</td>
<td>1978</td>
</tr>
<tr>
<td>24</td>
<td>as</td>
<td>Hong</td>
<td>1974</td>
</tr>
<tr>
<td>25</td>
<td>as</td>
<td>other</td>
<td>1972</td>
</tr>
<tr>
<td>26</td>
<td>as</td>
<td>University</td>
<td>1971</td>
</tr>
<tr>
<td>27</td>
<td>as</td>
<td>as</td>
<td>1962</td>
</tr>
<tr>
<td>28</td>
<td>Chan</td>
<td>Mr</td>
<td>1938</td>
</tr>
<tr>
<td>29</td>
<td>provided</td>
<td>shall</td>
<td>1921</td>
</tr>
<tr>
<td>30</td>
<td>measures</td>
<td>mitigation</td>
<td>1902</td>
</tr>
<tr>
<td>31</td>
<td>domestic</td>
<td>non</td>
<td>1894</td>
</tr>
<tr>
<td>32</td>
<td>control</td>
<td>system</td>
<td>1893</td>
</tr>
<tr>
<td>33</td>
<td>as</td>
<td>government</td>
<td>1890</td>
</tr>
<tr>
<td>34</td>
<td>as</td>
<td>Kong</td>
<td>1846</td>
</tr>
<tr>
<td>35</td>
<td>as</td>
<td>works</td>
<td>1832</td>
</tr>
<tr>
<td>36</td>
<td>as</td>
<td>government</td>
<td>1832</td>
</tr>
<tr>
<td>37</td>
<td>as</td>
<td>as</td>
<td>1823</td>
</tr>
<tr>
<td>38</td>
<td>as</td>
<td>monitoring</td>
<td>1798</td>
</tr>
<tr>
<td>39</td>
<td>civil</td>
<td>engineering</td>
<td>1783</td>
</tr>
<tr>
<td>40</td>
<td>as</td>
<td>building</td>
<td>1770</td>
</tr>
<tr>
<td>41</td>
<td>as</td>
<td>construction</td>
<td>1769</td>
</tr>
<tr>
<td>42</td>
<td>as</td>
<td>disposal</td>
<td>1768</td>
</tr>
<tr>
<td>43</td>
<td>as</td>
<td>engineering</td>
<td>1762</td>
</tr>
<tr>
<td>44</td>
<td>as</td>
<td>engineering</td>
<td>1758</td>
</tr>
<tr>
<td>45</td>
<td>as</td>
<td>management</td>
<td>1735</td>
</tr>
</tbody>
</table>

*Figure 7.2 Hong Kong Engineering Corpus ConCgram results*
Discussion

Which of these concgrams do you think it would be useful to focus on with an upper-intermediate level Engineering ESP group?

To illustrate this function, a concgram printout from the Hong Kong Financial Services Corpus (Warren 2010: 179) is reproduced below: this concgram looks at the relationship between the two words expenditure/increase (the concgram target words are always sequenced alphabetically).

![ConCgram: expenditure](image)

It is useful for our purposes that Warren (2010: 182) distinguishes between ‘lexically rich’ concgrams and other concgrams, explaining that a lexically rich concgram contains at least one lexical word in the combination. These would seem to be of greater potential ESP teaching value to us than concgrams with purely grammatical words.

Cheng (2010: 78) describes a case study of introducing students to ‘concgramming’, noting that ‘The students found the concgram analysis task challenging but learned a lot about phraseological patterns, particularly when the corpus texts are related to their disciplines’.

Corpora of Business English

While ‘Business English’ is a term that has been used for many years (and, of course, pre-dates modern corpora), it is worth pausing a moment to consider what
we mean by it. One obvious distinction is between spoken and written Business English: to illustrate how wide this division can be, consider Warren’s (2010) observation that in Hong Kong, while English is customarily used to write about business, Cantonese is typically used to talk about business. We could see this as a straightforward example of code-switching, but code-switching can be more complex in Business English as noted by Jenkins, Cogo and Dewey (2011) in the context of their discussion of English as a Lingua Franca (see Chapter 8):

... studies have revealed a complex relationship between English and other languages, especially in respect to multinational companies which increasingly use English as their corporate language. Although English is widely recognized as facilitating communication, there is also increasing evidence that the choice of language(s) used is a delicate issue ... and that other languages are also seen as important.

We need to keep in mind here Nelson’s (2006) distinction between English used to talk/write about business and English used to do business: Business English is not a monolithic phenomenon.

**Business English corpora and word frequency**

We begin our discussion of Business English with some observations about word frequency. Nelson (2006) carried out a key word analysis (see Introduction to this book) of Business English based on his 1.6 million corpus of spoken and written Business English: the BNC was used as a reference corpus for the purpose of establishing ‘keyness’. From this analysis, Nelson (http://users.utu.fi/micnel/business_english_lexis_site.htm) identifies 50 key words, which he divides into five semantic categories: people in business, business activities, business actions, business descriptions, and business events and entities. The categorisation is reproduced below:

*Table 7.4 Nelson’s key words in Business English*

<table>
<thead>
<tr>
<th>People in Business</th>
<th>Business Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>customer</td>
<td>business</td>
</tr>
<tr>
<td>manager</td>
<td>investment</td>
</tr>
<tr>
<td>supplier</td>
<td>delivery</td>
</tr>
<tr>
<td>distributor</td>
<td>payment</td>
</tr>
<tr>
<td>shareholder</td>
<td>development</td>
</tr>
<tr>
<td>employee</td>
<td>production</td>
</tr>
<tr>
<td>staff</td>
<td>communication</td>
</tr>
<tr>
<td>partner</td>
<td>competition</td>
</tr>
<tr>
<td>boss</td>
<td>takeover</td>
</tr>
<tr>
<td>management</td>
<td>distribution</td>
</tr>
<tr>
<td>Business Actions</td>
<td>Business Descriptions</td>
</tr>
<tr>
<td>------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>sell</td>
<td>high</td>
</tr>
<tr>
<td>manage</td>
<td>big</td>
</tr>
<tr>
<td>receive</td>
<td>low</td>
</tr>
<tr>
<td>confirm</td>
<td>global</td>
</tr>
<tr>
<td>provide</td>
<td>international</td>
</tr>
<tr>
<td>send</td>
<td>local</td>
</tr>
<tr>
<td>develop</td>
<td>competitive</td>
</tr>
<tr>
<td>discuss</td>
<td>corporate</td>
</tr>
<tr>
<td>achieve</td>
<td>strategic</td>
</tr>
<tr>
<td>improve</td>
<td>financial</td>
</tr>
</tbody>
</table>

**Business Events and Entities**

- sale
- merger
- trade
- package
- export
- service
- market
- earnings
- performance
- product

**Corpus Search**

2. Select Business Letter Corpus.
3. Enter one of the key words from Nelson’s list above.
4. From the data highlight what you would need to highlight for learners about how the word is used.

An unusual feature of Nelson’s ([http://users.utu.fi/micnel/business_english_lexis_site.htm](http://users.utu.fi/micnel/business_english_lexis_site.htm) research into business lexis) is that, as well as the key words above, he also identifies ‘negative’ key words, i.e. things business people do *not* normally talk or write about, identifying a number of non-business topics thus:

The analysis showed that Business English was semantically divorced from lexis concerned with personal issues, society, family, house and home and personal activities. Also used far less in business was lexis referring to distinctly negative states, and words used to express deep, reflective and emotive feelings were found to occur significantly less – the word *truth*, for example, was found to occur nine times less in Business English than in general English.
To remain with the theme of business vocabulary, the one million word spoken component of the CANBEC corpus provides the data for some interesting analysis of word frequency in spoken Business English by O’Keeffe, McCarthy and Carter (2007). Based on keyword analysis, using the conversation component of CANCODE as the benchmark corpus, O’Keeffe, McCarthy and Carter (2007) observe that we, us and our are all in the 50 most frequent words in CANBEC, whereas I and you are not. Even more strikingly, ‘None of the personal pronouns appear to be key in any way in ACAD [the academic component of CANCODE]’ (O’Keeffe, McCarthy and Carter 2007: 209). To take two further examples from O’Keeffe, McCarthy and Carter 2007: 209):

• While need appears at rank 9 in CANBEC, it is not a keyword in ACAD at all.
• CANBEC has so and problem in the top 50. ACAD has neither of these.

It is worth quoting in full the general observations about word frequency in spoken Business English made by O’Keeffe, McCarthy and Carter (2007: 210) because these observations remind us that the picture of language that emerges from corpus analysis is more important than discrete findings:

In the case of CANBEC, these key words and their contexts offer some insights into the interpersonal aspects of spoken business communication, and characterise it as (a) sharing properties with everyday informal and casual conversation, (b) sharing properties with institutional discourse (in this case academic) and (c) different from conversation and academic discourse, a special or unique register or genre which can be described by observing participants’ activity in the construction of relationships and identities, both individual and corporate, and the creation of business cultures …

Business English corpora and the frequency of collocations and chunks

We can now proceed ‘up the lexical scale’ from words to collocations and chunks. Mascull (2008–2013) gives a practical example of a materials writer working with a corpus in the field of Business English, noting how the Cambridge International Corpus gives him access to a wide variety of business communications including business sections of newspapers, business meetings and business lectures. He then illustrates how he moves from the corpus data to materials:

Take the word competition. The CIC lets me build, in a few seconds, a collocation table of 337 words that come in front of competition in the data. This would be pretty indigestible for even the keenest learner, but one of the roles of the ELT writer is to simplify and select, and in the context of corpora, this means selecting the most frequent and interesting features of the language,
therefore the items I included in a unit about competition in *Business Vocabulary in Use Advanced* were:

<table>
<thead>
<tr>
<th>cut-throat</th>
<th>ferocious</th>
<th>fierce</th>
<th>intense</th>
</tr>
</thead>
<tbody>
<tr>
<td>keen</td>
<td>low-key</td>
<td>stiff</td>
<td>tough</td>
</tr>
</tbody>
</table>

Nelson gives us an insight into frequency and lists the most common collocations/chunks in his corpus:

**Table 7.5** Nelson’s most common 2- and 3-word phrases in Business English

<table>
<thead>
<tr>
<th>2-word phrases</th>
<th>3-word phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>interest rates</td>
<td>a lot of</td>
</tr>
<tr>
<td>cash flow</td>
<td>one of the</td>
</tr>
<tr>
<td>market share</td>
<td>the end of</td>
</tr>
<tr>
<td>stock market</td>
<td>in order to</td>
</tr>
<tr>
<td>Wall Street</td>
<td>we need to</td>
</tr>
</tbody>
</table>

There is probably nothing surprising here: the 2-word phrases express core business content and the 3-word phrases reflect important business concerns: quantity, process, purpose and (on the face of it) obligation in the case of ‘we need to’. ‘We need to’ is, however, not quite as simple as it seems at first sight: it is also fourth in the list of 3-word chunks in CANCODE and, O’Keeffe, McCarthy and Carter (2007) argue, reflects a sense of corporate goals and/or acts as a kind of disguised imperative (i.e. the real meaning is, ‘You need to!’). In similar vein, Handford, a materials writer and researcher in the field of Business English, points out that the use of *have to* and *must* is distinctive in Business English:

> What we find instead [in Business English] is that *have to* is used 20 times more frequently than ‘must’ in business meetings, and that *must* is far more typically used in everyday situations … And when people in business do, rarely, use *must*, they are talking about what they themselves have to do, e.g. *I must do that*. It seems that saying *you must* threatens the other person’s self-image (or ‘face’) too much, and the same is true for UK and international situations.

**Business English and semantic prosody**

Discussion of business lexis gives us the opportunity to revisit briefly the topic of semantic prosody (see Chapter 2). Semantic prosody is the focus of Nelson’s (2006) study, who notes, for example, that while ‘boss’ collocates with negative adjectives, ‘manager’ does not display the same unilateral tendency. Much of this research, as Nelson (2006) notes, has to be qualitative: while the word ‘employee’ may not collocate particularly frequently with the word ‘benefits’, it collocates frequently with words which denote ‘benefits’, e.g. holiday entitlement, restaurant
Suppliers, by contrast, are typically referred to in terms of contractual obligations they have to fulfil (Nelson 2006: 224), e.g.

- The supplier shall deliver the Product within the territory …
- The supplier shall establish and maintain procedures …
- The supplier shall carry out a comprehensive system of planned …
- The supplier shall establish and maintain a documented quality …
- The supplier shall ensure that purchased product conforms to …

**Discussion**

2. Which of the activities on the webpage do you think are particularly useful?

**Learner corpora and ESP**

Research into Business English is a thriving area of ELF research (Jenkins, Cogo and Dewey 2011), though much of this research seems to focus on communication strategies such as accommodation strategies rather than lexico-grammatical analysis (Jenkins, Cogo and Dewey 2011). We now turn to a comparison of expert and ‘novice’ Business English carried out by Hewings (2012), based on two corpora he compiled for the purpose: a 120,000-word corpus of published research articles in the field of business administration and a 200,000-word corpus of student dissertations. One of the comparisons Hewings (2012) makes relates to the use of *it-clauses*: he notes that the students use a variety of strongly evaluative language in their writing to assert their claims, e.g.

- *it is amazing/strange/pointless/wise*
- *it is necessary/obvious/essential/vital/unacceptable/undeniable*

Such language, Hewings (2012) notes, was largely absent from the expert corpus and he concludes that ‘In the dissertations *it-clauses* of this type were mainly used in conclusions to present the implications for business practitioners of what students had found in their projects, while implications were expressed much less forcefully in the research articles’.

The use of the pronoun *I* in the expert and student corpora was also compared by Hewings. He found that *I* was used about four times more often in the student corpus than in the expert corpus, but also that it tended to be used in a different way: whereas the students tended to use *I* in opinion phrases, e.g. *I believe*, the experts tended to use *I* in a metadiscoursal way, e.g. *as I have pointed out*. 
We argued in Chapter 6 that learner corpora have a role to play in syllabus and materials design on the grounds that, while native speaker corpora can tell us what native speakers typically do, they cannot tell us what learners find difficult. The same argument applies in the EAP/ESP field: Flowerdew (2001, cited in Gilquin, Granger and Paquot 2007), for example, suggests that learner corpus EAP research can produce insights into important aspects of learner EAP writing, e.g. collocational patterning, pragmatic appropriacy, and discourse features. For their part, Gilquin, Granger and Paquot (2007: 327) refer to learners’ EAP problems in the following areas: frequency, register, positioning, semantics, and phraseology. A number of examples of such problems are provided by Gilquin, Granger and Paquot (2007: 327–329), including:

- Learners’ overuse of adverbs connoting a high degree of certainty (e.g. *absolutely*) and underuse of hedging adverbs (e.g. *apparently*).
- Learners’ tendency to overuse connectors in sentence-initial position (e.g. *however*).
- Misuse of phrases, e.g. the use of *on the contrary* when what is meant is ‘in contrast’.
- Malformation of chunks, e.g. *as a conclusion*.

**Corpus findings and the ESP syllabus**

In discussing ESP lexis we have been much concerned with lists, specifically the Academic Word List (Coxhead 2000), the Academic Collocation List (Ackermann and Chen 2013) and the Academic Formulas List. The use of frequency lists for lexical bundles to inform materials is also suggested by Swales. Frequency lists, Thompson (2006: 7) observes are ‘clearly of value both to syllabus writers and to materials developers’. But the question is, ‘How can they be of value?’ While the practice of learning decontextualised vocabulary lists has been somewhat rehabilitated in recent years (Folse 2007), it cannot be thought appropriate for the lists we have in mind for two reasons: first, we will be dealing here with a lot of abstract nouns rather than concrete terms for everyday objects and, second, learners, as we established in Chapter 2, need to know how a word operates in context and combines with other words. The limitations of list-learning are underlined by Coxhead (2000: 358):

One of the problems with developing a word list is that some learners and some teachers focus solely on working with the list alphabetically, for instance, starting at *abandon* and ending at *widespread*. Along the way, they meet strings of words that look similar, such as commence, comment, commission, and commit … Another problem is that students might never find the words in context in materials they are reading. They also might never practice the words in any meaningful way. For example, learners might focus only on the spelling and meaning of words, but not on using the words themselves in speaking and writing.
Part of the answer to the problems set out above may lie in how the lists are treated. In relation to collocations, for example, Ackermann and Chen (2013) suggest that collocations can be grouped in a number of ways, e.g. adjective and noun combinations, common frames or collocation families. Ackermann and Chen (2013: 241) also argue that the ACL may not be as daunting as it first looks if we focus on ‘recurrent frames’:

The implication for EAP pedagogy at the lexico-grammatical level is that despite a seemingly large number of collocations in the ACL, the actual teaching and learning load should be manageable as many of the words in the ACL are often part of a ‘recurrent frame’, and understanding the frame will give learners a sense of familiarity when encountering new collocations …

As an example of a ‘recurrent frame’ they give (be) highly + past participle, e.g. be highly regarded, be highly respected, be highly motivated etc.

With pedagogic concerns in mind Simpson-Vlach and Ellis (2010), as we noted earlier, suggest how their Academic Formulas List can be categorised, but, as we saw, their categorisation is quite complex and at a level of abstraction that would not, I would argue, make it easy for teachers and materials writers to translate it into a syllabus: what is systematic linguistically is not necessarily helpful pedagogically. It is fair to say, however, that Simpson-Vlach and Ellis (2010) see their work as a starting point rather than the final word:

It is our hope that this functional categorization, along with the FTW rank-ordered lists, will facilitate the inclusion of AFL formulas into EAP curricula, and that further work on the pedagogical value of the AFL will take these results as a starting point.

Frequency and classification move us nearer towards a syllabus, but, as well as frequency, we need to take into account coverage, range and, crucially, learnability. Ward’s (2009) engineering word list, for example, does take range into account by ensuring that words included are frequent in all five engineering sub-disciplines included in the corpus. A further limitation of frequency as a determining principle for an ESP lexical syllabus is that we need to account for what particular sets of learners find difficult (Gilquin, Granger and Paquot 2007), i.e. a learner corpus perspective is useful.

It is difficult to maintain an absolute division between syllabus and methodology and it may be that the answer to the syllabus question lies, at least to some extent, in the methodology we adopt. What is required, I would argue, is a kind of ‘informed noticing’, as appears to be suggested by Swales:

… encouraging learners to notice these multi-word units through repeated exposure and through activities such as matching and item identification. Consciousness-raising tasks which offer opportunities to retrieve, use and
manipulate items can be productive, as can activities which require learners to produce the items in their extended.

Awareness-raising is also proposed as a goal by Ackermann and Chen (2013: 245), though they recommend a combination of explicit teaching and consciousness-raising work.

Explicit teaching needs to be complemented by providing students with the opportunity to encounter collocations when dealing with academic texts. Combining both explicit and implicit teaching will enhance learners’ receptive and productive collocational knowledge and thus improve their academic English proficiency.

Through these proposals, it can be seen that the frequency lists we have referred to can act as points of reference for teachers, materials writers and students in developing a feel for what kind of lexis to look for in, for example, the written and spoken texts used in the classroom. The process of noticing important academic words, collocations and chunks will probably start as guided classroom activity with the ultimate aim of helping learners to become autonomous in this respect. Online tools such as the Academic Word Highlighter and wordandphrase.info are ideally suited to this kind of autonomous work.

**English for Specific Purposes corpora and DDL**

In the previous chapter, we noted the argument that DDL might be better suited to higher level analytical learners. There is a case, then, that DDL will suit our EAP (and ESP) learners, though even in the Higher Education context it is not necessarily common practice: Lee and Swales (2006: 69) report a study by Jarvis (2004) that found that concordancing was being used, on average, in about one in ten of the ESP/EAP courses in the UK. The situation may have changed since then, of course, but my personal impression is that it still has ‘minority sport’ status in UK Higher Education. Two relevant case studies of DDL in an EAP context are reported by Lee and Swales (2006: 70), which I summarise below:

- Yoon and Hirvela (2004) found that DDL helped their students with writing skills, and writing confidence. It seemed, however, to be less helpful in helping them with word meanings, with academic reading and with exam preparation for TOEFL.
- Coniam (2004) used his publications, and a colleague’s publications, as a kind of pedagogic EAP corpus for DDL and concluded that this approach could be helpful in allowing students to compare their writing with established writers in the field.

Lee and Swales (2006) also report on their own extended experiment with DDL. The main stages of their experiment are summarised below:
PhD candidates or advanced pre-PhD candidates were selected for the experiment on the grounds that they had a ‘vested interest in fine-tuning their knowledge of academic writing and speaking’ (Lee and Swales 2006: 59).

Participants were given an induction period during which they were introduced to concordancing tools and given practice in corpus analysis.

Participants gradually compiled two corpora themselves: one corpus comprised their own writing (e.g. draft course work or draft journal articles); they compiled the second corpus from journal papers in their own field.

Participants used the techniques learnt in step 1 to investigate the two corpora and to make comparisons between their own writing and established expert writing in the field.

Lee and Swales (2006) report that, while, on reflection, they would make some changes to the preparatory stages, the general response from the participants was positive as they found the process empowering. The advantage of such an approach, Lee and Swales (2006: 71) argue, is that it is ‘decentering’:

We believe the corpus approach is decentering because: (i) it allows non-native speakers a chance to make their own discoveries about what is done in the language, instead of relying on native-speaker intuitions or grammar/style books; (ii) it typically involves texts from a variety of different writers/speakers (not all of whom are necessarily native speakers), instead of just one native-speaker teacher standing at the front of the classroom. In other words, it is decentered away from the native-speaker and away from any one individual person or grammar book or stylistic convention.

It is clear from this experiment that a great deal of autonomy was required on the part of the students, and we saw in the previous chapter that increased autonomy has been proposed as a benefit of DDL. The need for autonomy can, however, be one of the challenges, or potential pitfalls. It is to these challenges of DDL in an EAP/ESP context that we now turn.

The most obvious pitfall we identified in relation to DDL in the previous chapter was the risk of ‘drowning in data’ when faced with hundreds of concordance lines, many of which may contain obscure lexis or cultural references. Here we must recall that we can be selective in the concordance lines we present, omitting those we consider inaccessible. A further potential objection to DDL in the EAP classroom is the pedagogic cost–benefit analysis: focusing on the behaviour of even a single word through concordance discovery activities can be a very time-consuming activity. In this respect, I would suggest that we see classroom DDL activity as learner training to prepare learners to carry out this kind of work in a self-directed, autonomous way. We must also consider the kind of language work that can be carried out through concordance work. As Lee and Swales (2006: 570) point out ‘… a wholesale commitment to a corpus-based approach may not be fully effective, since concordancing tends to work better at
the lexico-grammatical and phraseological levels rather than the structural level’. This is true: concordance work tends to focus on the micro-level rather than the macro-level. We need to keep in mind, however, that there is no need for a ‘wholesale commitment’ to DDL: there is no technique in language teaching (as far as I know) which suits all learners all the time for all kinds of language work. Finally, there is the logistical problem of finding corpora suitably specialised for our learners: don’t forget the ‘quick and dirty’ option! Thus far we have focused mostly on ESP corpora as resources for syllabus designers, materials writers and learners. It is worth noting, then, Hewings’ (2012: 14) observation that corpora can be useful resources for teachers in their everyday ESP work:

DDL views corpora primarily as resources for students to learn from. They can also, of course, be resources for teachers to explore, to check that their intuitions are supported by the evidence found in a large data set, to make new discoveries about how language is used, and as a source of authentic examples.

Corpus-related language work does not, of course, necessarily involve concordance printouts. Crawford Camiciottoli (2010: 103) makes a number of specific proposals for the exploitation of a small corpus of business lectures. Among her suggestions are:

- Excerpts from the lectures could be used to focus on key specialised vocabulary (identified in the corpus) through techniques such as gap-filling, matching or multiple choice.
- Students could be given a list of key specialised vocabulary (from the corpus) before the lecture as a basis for a prediction task.
- Students could be given different lists of specialised vocabulary and asked to choose the one which best fits the lecture.
- Students could be asked to note down specialised lexis and then compare it with a list prepared by the teacher.
- As a follow-up task, students could prepare a short oral or written summary that incorporates the items that they identified.

A focus on key lexis is also suggested by Scott (2005), who argues that key words can be used as the basis for a number of activities, including:

- Predicting the content of a listening/reading text.
- Suggesting other text types which might contain the same lexis.
- Generating a text from key words.
- Listening and taking notes using key words.
- Generating an oral presentation from key words.

We should also note that Scott suggests practising the pronunciation of key words as pronunciation does not seem to receive much attention in ESP/EAP work.
The use of corpora in EAP is far more common in reference works such as dictionaries than it is in coursebooks, though native speaker corpora wield far more influence in this field than learner corpora (Gilquin, Granger and Paquot 2007 – see also Chapter 6). Gilquin, Granger and Paquot (2007) argue that learner corpora should play a far greater role in pedagogic reference works:

What L2 learners need is EAP resource books addressing the specific problems they encounter as non-native writers. By showing in context the types of errors learners make, as well as the items they tend to underuse or overuse, learner corpora make such an approach possible. Yet, hardly any materials writers up to now have taken up the challenge of using learner corpus data.

A corpus ‘product’ which does seem to have had some influence on both reference works and coursebooks is the Academic Word List. Coxhead (2011) lists a number of resources which exploit the Academic Word List, among which are:

- Contemporary Topics (Longman)
- Essential Academic Vocabulary: Mastering the Complete Academic Word List (Houghton Mifflin)
- English for Academic Success Series (Houghton Mifflin)
- Improve Your English Vocabulary (Dee Publications)
- Academic Word Power series (Heinle ELT)
- Focus on Vocabulary: Mastering the Academic Word List (Longman)
- Academic Vocabulary in Use (Cambridge University Press).

The classroom ideas we have focused on above all focus heavily on lexis, so it will be useful to look at an example that is more broadly focused on the writing skill. Adel (2010) describes an activity using the MICUSP corpus of student writing. The task set for the novice writers was (Adel 2010: 43): ‘What do academic writers say … (a) when they give an example, (b) refer to other texts or researchers (c) introduce the topic (d) start their conclusion section?’ Among the points noted by the novice writers through this task, Adel (2010: 43) observes, was how writers established the ‘urgency factor’ (i.e. ways of establishing the importance of the topic) in the introduction using words such as critical, fundamental, central, key.

Two other practical points about this experiment are worth noting: 1) after the work in class, students were sent slides summarising the answers they had found; and 2) the corpus classroom work was harmonised with the work of the regular writing teachers. These two features are important, I would argue, in ensuring that learners see some tangible outcome to their activity and see the corpus work as integrated with the rest of their course. The outcome of the experiment was, Adel (2010: 43) reports, ‘primarily positive’.
Finally in this chapter, we review some possibilities for corpus-based EAP materials. Quite detailed suggestions for developing EAP materials from MICASE are available at http://www.press.umich.edu/pdf/0472030728-part7.pdf

A feature of the MICASE approach to corpus-based EAP materials is that, while lexico-grammatical work is included, it does not dominate, and other areas of focus, facilitated by the search interface, are suggested and exemplified:

**Situational**

1. Target a speech event, context, or setting.
2. Target particular speaker characteristics.
3. Functional-Pragmatic.

   1. Target a particular pragmatic function. For example, look up ‘definitions’ or ‘disagreement’ or ‘advice’ in the Pragmatic Features Overview Chart (pp. 70–81) for transcripts with this type of content. Use the Speech Event Abstracts (pp. 101–252) to select transcripts to investigate further.

   2. Target certain discourse style characteristics. For example, in linguistic terms, what seems to stimulate interactivity in a class?

   3. Contrast functional-pragmatic language across fields or disciplines. For example, use the Pragmatic Features Overview Chart (pp. 70–81) and/or the Keyword Index (pp. 253–265) to identify transcripts with particular features (e.g. question-and-answer periods) or conversation topics (e.g. experiments) to compare discourse structure, register, vocabulary, etc.

There are also specific suggestions as to how the corpus transcripts can be ‘cleaned up’ for use in the classroom:

- Number the lines or turns.
- Capitalise (i’s, and first word in each turn, etc.).
- Change or reformat some transcription conventions (overlapped speech, laughter, pauses, etc.).
- Adjust the format (line spacing, indents, including hanging indents for each speaker, etc.).
- Assign names to speakers. (The header of the transcript can be checked to determine the speaker’s gender, if desired.)
- Delete some dysfluencies, particularly if they are excessive or unduly intrusive (i.e., *uhs* and *ums*, false starts, repetitions, etc.).
- Correct the grammar. Even though the beauty of MICASE is its authenticity, sometimes it just seems wrong to include a grammatically incorrect utterance in teaching materials. Consider the pros and cons.
Materials generated from BAWE are available at http://learnenglish.britishcouncil.org/en/writing-purpose/writing-purpose

There are also tips here on how to search http://www.press.umich.edu/pdf/0472030728-part7.pdf

BAWE (British Academic Written English Corpus)

As a further example of the step from corpus to materials, we can consider Mudraya’s (2006) proposal for combining DDL with a lexical approach in the context of teaching engineering English. An example activity, based on a concordance of the word solution is reproduced below. This example shows, I would argue, how a concordance of a keyword can generate a variety of useful activities.

Figure 7.4 Concordance: solution

Corpus Search

The final example is my own and focuses on identifying key academic words and their behaviour. It also involves using the following website: http://www.wordandphrase.info/analyzeText.asp
Students find an extract from an article they have read and found interesting recently – the abstract, the introduction or the conclusion of the article would be suitable (as these often include the key words for the particular topic).

Go to http://www.wordandphrase.info/analyzeText.asp

An empty text box will appear.

If the article you have chosen is available electronically, copy and paste the abstract into the text box. If it is not available electronically, you will need to type it in.

Click Search.

Your text will appear in an adjacent box with words colour-coded for frequency.

Click the red ACAD symbol and frequent academic words will be highlighted.

Click on any word that interest you and you will get a list of frequent collocations and a lot of concordance examples.

Conclusion

In conclusion, we have seen that ESP has probably been more influenced by corpus linguistics than has general English. Even in the field of ESP, however, this influence is restricted and inconsistent: it is most evident in EAP and more evident in materials and syllabuses than it is in classroom practice. Thompson (2006: 15) offers a number of suggestions as to how the relationship between corpora and EAP might develop (I would argue that the suggestions apply to ESP in general).

In order to develop the direct use of corpora in EAP teaching and course design, it is proposed that the following are required:

- empirical research into the usefulness (or not) of corpora in direct teaching
- improved channels for dissemination of practice
- greater access to relevant corpora (MICASE/BASE, learner corpora, corpora of relevant exemplar texts) and corpus analysis tools, and an increased number of corpora created for use by practitioners rather than primarily for researchers
- thorough training in corpus analysis techniques, and the evaluation of such methods, into EAP teacher development courses, especially within MA TESOL programmes.
References


Introduction

It might at first sight seem strange to have a chapter that looks at the limitations of corpora, and reservations about corpora, in a book that aims to promote the use of corpora in ELT. However, my belief is that we will only be able to make optimum use of corpora if we have a measured view of their value and a strong sense of their most appropriate applications in ELT. After all, nobody would dispute that a hammer is a valuable tool, but it helps to know that it is good for hitting nails rather than cutting down trees. In this chapter we deal first with general reservations about the nature of corpus evidence before dealing in some detail with the ELF school of thought that has suggested that ELT corpora have tended to privilege the native speaker.

Limitations of corpora

Hunston (2002: 22–23) notes four main limitations of corpora which I summarise and expand upon below:

1. Corpora will tell you whether something has occurred and whether it is frequent, but it cannot tell us what is possible in a language. At any moment, a speaker can produce a form or a turn of phrase which follows the rules of the language but is not captured by a corpus. In other words, corpora capture what people do with a language, not what they know about it. This is shown clearly in the case of idioms: while corpora may show that because speakers assume that interlocutors share knowledge of the idiom they may only use it elliptically, e.g. *oh well, a stitch in time* … in almost all cases they will obviously know what the whole idiom is. While the Chomskyan competence–performance distinction is controversial, this does seem to be an example of how competence (in this case knowledge of the idiom) can differ from performance (the tendency to produce the idiom elliptically).

2. Corpora, no matter how rigorously designed in terms of demographic and generic balance can never be truly representative of a language. While corpora
grow ever larger, the point made by Gavioli and Aston (2001: 238) holds
good: ‘The largest corpora of English are still smaller than the average adult
user’s experience of the language, and very different in their composition…’

Corpora can give us plenty of evidence in the way of examples, but they do
not provide us with interpretations. Interpretations must come from qualitative
analysis by the corpus researcher, a process which can be time-consuming
and imperfect.

Corpora cannot capture the whole context – visual, spatial or social – in
which the language was used. And in relation to spoken language, it cannot
(yet) capture all the accompanying prosodic and paralinguistic features. We
should add here that progress has been made with multi-modal corpora since
Hunston (2002) made these observations, but it is difficult to believe that
corpora will ever capture the totality of the communicative experience,
especially speakers’ intentions at any given moment. As Cook (1998: 58)
points out: ‘It is a truism to observe that there is no straightforward correlation
between the words people use, the intentions they had in them, and the
interpretations other people put on them’. Language, as Cook (1998: 57)
argues, is too complex a phenomenon to be captured by records of use: ‘If the
traditional concern of linguistics – language in all its cultural and psychological
complexity – could be replaced by a neat computer bank of data, life would
be much simpler’.

Corpora, the native speaker and English as a
Lingua Franca

Native speakers in perspective

Thus far, with the exception of our discussion of learner corpora, we have largely
confined our discussion to the findings of native speaker corpora. In relation to
teaching, I have, I trust, been careful to speak of the potential value of such
findings for ELT. Now is a convenient time to justify this hedging on my part and
ask: why are these findings only of potential value? What reservations might there
be about the influence of native speaker corpora on ELT? One reason is that the
relevance of corpus findings from native speaker corpora, and, in particular,
native speaker spoken corpora, has been questioned on sociocultural grounds. The
main question in relation to the sociocultural debate is expressed by Timmis
(2013: 84) thus:

If the majority of communication in English takes place between non-native
speakers, and that preponderance is highly likely to rise, why encumber
learners with the ‘luxury’ items of native speaker spoken language? And if
the primary function of these features is the construction and negotiation of
relationships and identity, does teaching these features involve imposing a
false identity on the learners?
The same question is posed, I would concede, more elegantly and succinctly by Prodromou (1998: 266): ‘What is real for the native speaker may also be real, say, for the learner studying in Britain, but it may be unreal for the EFL learner in Greece and surreal for the ESL learner in Calcutta’.

Given that native speakers are now clearly in the minority in the English-speaking world, what authority should native speaker corpora hold over English language users who have neither the opportunity nor desire to interact with native speakers? It is a debate that has generated much discussion, but I would like to illustrate how it can raise small but concrete questions at classroom level. Some of my students, for example, speak about ‘public schools’ in their country: now, a search of a British native speaker corpus would reveal that in British English the term ‘public school’ actually refers to private, fee-paying schools (often boarding schools) such as Eton and Harrow. The British use of ‘public school’ is deeply embedded in its history as an empire and its social structure. To take another example, in the context of our discussion of collocations, I noted that my learners frequent say, ‘I’m going to study a masters’, while a native speaker would typically say, ‘I’m going to do a masters’. A reasonable question would be, ‘What right (or need) do I have to correct study a masters when it is perfectly comprehensible and perhaps the norm for more people than it is to say do a masters”? While all language use is culturally determined to some extent, some cases are more obvious than others. If we return to our discussion of reported speech we can ask, for example, whether we should continue to teach the (native speaker) canonical rules of reported speech which, we have noted, are rarely applied by native speakers in conversation, or, alternatively, introduce learners to the new spoken quotatives such as ‘I was like’. This form is easier to manipulate than the back-shift rules of reported speech, but it is also socially marked in the corpora suggesting it is frequent in the conversation of (younger) native speakers, often under the age of 30 (Adolphs and Carter 2003). If we choose not to impose native speaker grammar all the time, how far can we go? Willis, for example, questioned whether it is worth correcting an utterance such as ‘he suggested me to go’ (I hope it isn’t – I’m tired of correcting it) – Willis argued that such an utterance is perfectly comprehensible and it is grammatical enough, i.e. it follows the basic rule of other reporting verbs such as tell/ask/advise/encourage someone to do something. However, if a learner chooses to extend the rule and produced utterances such as she accused me to do it or she blamed me to do it or he congratulated me to do it, will these utterances also be generally comprehensible?

**English as a Lingua Franca research**

One response to this debate can be seen in the growing body of research (Seidlhofer 2001; 2004; Jenkins, Cogo and Dewey 2011; Cogo and Dewey 2013) into the use of English as a Lingua Franca (ELF). We need first to consider the definition of ELF: Jenkins, Cogo and Dewey (2011) draw attention to the definition on the VOICE website, i.e. ‘an additionally acquired language system which serves as a
common means of communication for speakers of different first languages’. A key feature of this definition, which Jenkins, Cogo and Dewey (2011) emphasise, is that it does not exclude native speakers. However, it does, they argue, implicitly that native speakers will need to learn to use this system: they are not native speakers of ELF (though they acknowledge that native speakers may have a head start in acquiring this additional language system).

The basis of much ELF research to date has been the one million word VOICE corpus (Vienna–Oxford International Corpus of English), though there is now a one million word ACE (Asian Corpus of English) discussed (see page 000) by Kirkpatrick (2014). As VOICE is a groundbreaking ELF corpus, it is worth considering the description and rationale provided on the website to contextualise the discussion of ELF findings that follows in this section: http://www.univie.ac.at/voice/

In the early twenty-first century, English in the world finds itself in an ‘unstable equilibrium’. On the one hand, the majority of the world’s English users are not native speakers of the language, but use it as an additional language, as a convenient means for communicative interactions which cannot be conducted in their mother tongues. On the other hand, linguistic descriptions have as yet predominantly been focusing on English as it is spoken and written by its native speakers.

The Vienna–Oxford International Corpus of English seeks to redress the balance by providing a sizeable, computer-readable corpus of English as it is spoken by this non-native speaking majority of users in different contexts. These speakers use English successfully on a daily basis all over the world, in their personal, professional or academic lives. We therefore see them primarily not as language learners but as language users in their own right. It is therefore clearly worth finding out just how they use the language. This is exactly what VOICE seeks to make possible.

There are three interesting questions which emerge immediately:

1. How do you define a ‘successful user’ of English? While ‘native speaker’ is sometimes a slippery term, or at least portrayed as such by argumentative academics, ‘successful user’ does not seem to be any easier to define objectively as a term.
2. The statement ‘We therefore see them [speakers of ELF] primarily not as language learners but as language users in their own right’ (my italics) is an interesting hedge. What exactly is the distinction between a successful user and an advanced learner?
3. If ELF is an additionally acquired system (my italics) does it qualify as a variety of English?

We have already referred to ELF research in relation to Business English and EAP so we now consider examples in relation to general English. With reference to lexi-co-grammatical research, Jenkins, Cogo and Dewey (2011) point out that
variants need to satisfy three criteria before they can be regarded as characteristic of ELF users. ELF variants must be:

1. systematic in nature as established by qualitative and quantitative analysis
2. frequent and widespread, occurring across numerous speakers and linguacultural backgrounds
3. communicatively effective.

Examples of ELF variants which meet such criteria are (Seidlhofer 2004: 220):

- third person present tense without ‘s’
- interchangeable use of the relative pronouns ‘who’ and ‘which’
- omission of definite and indefinite articles where they are obligatory in ENL, and inserting them where they do not occur in ENL
- use of a narrow range of tag questions (e.g. isn’t it? or no? instead of shouldn’t they?)
- use of ‘redundant’ prepositions
- more frequent use than in native speaker varieties of certain verbs of high semantic generality, such as do, have, make, put, take
  - use of infinitive-constructions where that-clauses are required in native speaker varieties, e.g. I want that
- explicitness not normally found in native speaker varieties (e.g. black colour rather than just black).

What will be immediately obvious to experienced English language teachers is that these features are often seen as fossilised errors, as indeed Jenkins, Cogo and Dewey (2011: 290) acknowledge:

> Many of these features are precisely the language forms that tend to be treated as ‘fossilized’ in ELT, that is, they would be seen as evidence of incomplete learning, errors which are resistant to the NS-based corrections of language teachers.

**Discussion**

How far do you see the examples above as errors or as ELF variants?

Much lexical work in the field of ELF research focuses on creativity. Jenkins, Cogo and Dewey (2011) refer, for example, to the work of Bjorkman (2009) on morphological innovation, citing the examples of *discriminization* and *levelize*. Jenkins, Cogo and Dewey (2011: 290–291 also cite the work of Pitzl (2009) on idiomaticity:
Pitzl shows that idioms in ELF can vary formally from their ENL equivalents, but without inhibiting (and possibly even enhancing) their functionality. The example of ELF idiomaticity provided in the title of the paper, ‘we should not wake up any dogs’, is an elegant illustration of what she means.

It is most interesting here that Jenkins, Cogo and Dewey (2011) argue that (native speaker) institutionalised forms constrain naturally-occurring processes of language production, as one could take exactly the opposite stance and argue that a readymade stock of collocations and chunks enhances fluency and mutual comprehension. We argued this in Chapter 2, noting that it is more difficult to decode and encode word-by-word than it is to access collocations and chunks: the need to coin collocations and idioms afresh slows the speaker down and makes comprehension more difficult for the interlocutor. There has been a good deal of discussion in ELF research of accommodation strategies, but creating idioms afresh does not seem to qualify as ‘accommodating’ the interlocutor (though fresh idioms may enliven the exchange). This might be a case of what Seidlhofer (2001) calls ‘unilateral idiomaticity’, i.e. the speaker uses a form not comprehensible to the interlocutor.

The example of we should not wake up any dogs also raises the question of the nature of creativity, a quality of ELF interaction which is often stressed in ELF research. It could, however, be argued that the utterance we should not wake up any dogs is simply resourceful (or even more simply translation) and that real creativity requires a knowledge of conventions and judicious breaking of these conventions for deliberate effect. Ultimately, whether one sees here resourcefulness or creativity is an ideological question, because the conventions to be subverted in pursuit of creativity are those of the native speaker.

A further aspect of the ELF debate we need to consider is whether there is a qualitative difference between native speaker varieties and lingua franca varieties (of any language). Cook (2002), for example, pointed out that a native speaker language (regardless of the number of speakers) has to fulfil all the communicative functions of its community of speakers, e.g. ceremonial, ritual, legal, intimate, artistic, ludic or even abusive functions. A lingua franca, because of its typical contexts of use, is not likely to be called upon as often for these purposes as a native speaker variety. While learners may or may not need to exercise the range of communicative functions in English mentioned above, they may well be interested in them out of natural curiosity. To treat English merely as a functional lingua franca may lead to an excessively utilitarian approach to ELT.

The position of ELF researchers with regard to the relevance of ELF research to ELT has not always been clear, perhaps in part due to what ELF researchers see as misrepresentation of their position by vituperative critics (e.g. Sobkowiak 2008). It is as well then to quote the position of ELF in relation to ELT set out by Jenkins, Cogo and Dewey (2011: 306):

ELF research, then, is not about determining what should or should not be taught in the language classroom. Rather, ELF researchers feel their
responsibility is to make current research findings accessible in a way that enables teachers to reconsider their beliefs and practices and make informed decisions about the significance of ELF for their own individual teaching contexts.

It seems appropriate, then, to end this section with the aims of the ACE corpus (Asian Corpus of English), to which we referred briefly earlier in this chapter, as these aims build on the approach of VOICE researchers and encapsulate the kinds of ELF research which have been carried out and the areas in which it might develop. The ACE corpus can be accessed at http://corpus.ied.edu.hk/ace/

**Ten objectives of the ACE project**

http://ec-concord.ied.edu.hk/ace/index.php?m=news&a=index

1. What, if anything, notwithstanding all the diversity, might emerge as common features of Asian ELF use?
2. What seem to be the most relied upon and successfully employed grammatical constructions?
3. What are the factors (pragmatic) that lead to misunderstandings and communication breakdowns on the one hand or communicative success on the other?
4. Is the degree of approximation to a L1 variety of English always proportional to communicative success?
5. Are there commonly-used constructions, lexical items and sound patterns that are ungrammatical in standard L1 English but generally unproblematic in Asian ELF communication?
6. What strategies do speakers employ when negotiating meaning?
7. What role, if any, might the L1 play in the creation of distinctive features?
8. What similarities and differences can be identified between European and Asian ELF?
9. Do the findings in point 8 suggest any universal but distinctive features/grammar of ELF?
10. Can any of the universal but distinctive features be explained by the motivations of syntactic simplification and regularisation that have been shown to be at work in ‘traditional Englishes’ for centuries?

**Classroom options for models of English**

We have seen from the arguments and examples above that native speaker corpora have played an important role in the debates as to what model of English it is appropriate to offer in our classrooms given current and predicted patterns of use of English worldwide. One response would be to abandon native speaker norms as both an unattainable and inappropriate target. We see an example of such a response from a materials writer below (cited in Timmis 2003):
I teach and write books based on International English. I have no corpus to tell me what this is, but I help my students to be simple, clear and direct, nothing more and nothing less. If I find what they say difficult to understand (or if other students do) I give them feedback on their errors to make them easier to understand. If they can communicate their meaning accurately although the form/lexis is not native-speaker like I’m not too worried.

This remark raises a number of questions relevant to the debate about the role of corpora in contemporary ELT:

• How does this native speaker materials writer know what international English is in the absence of a corpus?
• Is it always appropriate to be ‘direct’ in communication of any kind?
• Are the learners themselves content to be ‘simple, clear and direct’ whatever their overt purposes are for learning English?
• How can you judge what an error is, and how can you correct an error, in the absence of an explicit set of grammatical norms?

I would argue that this materials writer’s response to the debate about the position of native speaker English in ELT was simplistic. Willis (1999) presented the concrete options available to the English language teacher in the light of the debate we have outlined. A number of options for models of English in the classroom are proposed by Willis (1999), which are summarised below along with my commentary. We should note first that the term ‘International English’ seems to have disappeared from the debate in recent years to be replaced by English as a Lingua Franca (ELF – see below).

Option 1: Teach Standard British English

Willis (1999) considers this option to be both ‘undesirable and impractical’: undesirable on sociocultural grounds and impractical on pedagogic grounds. It could be countered, however, that the question of what is socioculturally appropriate ultimately rests with the learner. In practical terms, there are already extensive, but not, of course, complete grammatical descriptions of British English, which corpus evidence is helping us to refine. Native speaker norms represent, then, a definable target for students albeit one which they almost certainly realise they will never attain: rightly or wrongly, they recognise such a target as ‘authentic’. The same arguments would, of course, apply to any native variety of English.

Option 2: Define a form of ‘International English’ and teach that

The first well-publicised step in this direction was Jenkins’ (1998) work on a Lingua Franca Core (LFC) for pronunciation, an attempt to identify those sounds
which seemed central to ensuring intelligibility. Willis (1999) suggested that it may be useful to research a grammatical Lingua Franca Core. The International English option is also supported by Modiano (2001: 344):

The teaching and learning of a geographically, politically and culturally ‘neutral’ form of English, which is perceived as a language of wider communication and not as the possession of the native speaker, is one of the few options we have at hand if we want to continue to promote English language learning while at the same time attempting to ‘neutralize’ the impact which the spread of English has on the cultural integrity of the learner.

Widdowson (1994: 385) suggests that given the desire to be mutually intelligible and the global communication network, International English will tend to ‘stabilise’.

There are, however, two problems, or at least perceived problems with the LFC argument:

1. It presents a formidable corpus research challenge considering the huge number of English users in the world and their geographic, cultural and linguistic diversity.
2. It could lead to a rather minimalist approach to the language syllabus, promoting the transactional at the expense of the affective.

**Option 3: Offer a range of Englishes in the Classroom**

It is undoubtedly important for learners to be aware that different Englishes exist and to approach the task of communicating with people who speak a different variety without prejudice, but there are problems with exposing learners to different Englishes: it could be confusing and frustrating, and there is the question of which varieties to select. It would be highly impractical to expose learners to all the varieties available. While one occasionally hears non-native voices in mainstream ELT coursebooks, it always seems to be at a rather token level, and publishers report a strong preference on the part of both teachers and learners for native speaker voices.

**Option 4: Offer successful L1 speakers as models**

The notion here is that a group of French learners of English could be offered successful French learners of English as models. Alptekin (2002: 63) also argues that ‘successful bilinguals’ would make good models, especially because of their ‘intercultural insights and knowledge’. It can be argued, however, that at least part of learning any language should involve exposure to different cultures and different language communities – this option would also clearly be restricted to monolingual classrooms.
Option 5: Give learners exposure to native speaker English but adopt a consciousness-raising (C-R) methodology which places a very low premium on conformity

As this is clearly Willis’ (1999) favoured option I will quote him in full on this one:

See language as a meaning system and encourage learners to develop their own systems. Carry out C-R [consciousness-raising] work which encourages learners to focus on form, but place a very low premium on testing for conformity. Cut out the focus on forms which have little communicative value (e.g. question tags). Look for productive generalisations (e.g. V+N+Infin) and pay much less attention to exceptions like suggest.

The potential problem here is that it may appear that the teacher is imposing a low premium on conformity to native speaker norms and to be running the risk of offering an impoverished syllabus. The argument that question tags have ‘little communicative value’ is an interesting one. Clearly, they add nothing to propositional content, but they may well work affectively as an attempt to engage the listeners. Here it is worth noting Willis’ (1999) distinction between ‘accuracy’ and ‘conformity’: he considers ‘accuracy’ to be the ability to operate a stable grammatical system comprehensibly and ‘conformity’ to mean adherence to native speaker grammatical rules. There is a danger here of sliding into a purely utilitarian approach to language learning, as Carter (1998: 51) remarks:

Learners should not be patronised by being told that they do not need to bother with all this real English. They should not be disempowered and syllabuses should not be deliberately impoverished. Also, learning a language should, in part at least, involve developing something of a feel for that language.

Option 6: Include the study of language and dialects in a language teaching programme

This option seems to be a potential adjunct to the other five options rather than an alternative option. There would indeed seem to be a strong argument for preparing learners to understand the diversity they may encounter when speaking English and to take a positive view of the challenge.

Discussion

Which of the options above do you think is preferable in your teaching/learning context?
It was in the context of the debate about the relevance of native speaker corpora in contemporary ELT that Timmis (2002; 2003) carried out a survey of learners’ and teachers’ attitudes to the question of aiming for native speaker norms in general, and native speaker spoken norms in particular. The main variables Timmis (2002) wanted to investigate in relation to learners are below:

1. Learning context: The learning context was categorised following Kachru’s model as inner circle, outer circle or expanding circle.
2. Current patterns of use of English, i.e. whether respondents used English mainly with native speakers or other non-native speakers.
3. Predicted patterns of use, i.e. whether respondents anticipated using English more with native speakers or other non-native speakers in the near future.

There were around 400 responses to the student questionnaire from adult learners of intermediate level and above in 15 different countries and, while fully acknowledging the limitations of such research, Timmis (2002) summarised the findings thus:

- A clear majority of learners in inner and expanding circles aspired to native speaker norms, whether or not they envisaged using English with native speakers in the future.
- The majority in the outer circle countries (India, South Africa and Pakistan) preferred local norms.
- A clear majority of learners appeared to aspire to native speaker spoken norms, but were far less certain when shown an example of authentic native speaker speech.

The main variable to be investigated in the teacher questionnaire was the native/non-native variable. This questionnaire drew around 250 responses from teachers in about 50 countries, fairly evenly divided between native speaker teachers and non-native speakers. Probably the clearest result in this case was that many teachers felt the decision was best left to the learners. There was some indication, however, that non-native teachers felt a greater attachment to native speaker norms than native speaker teachers.

While, as Timmis (2012) acknowledges, his survey only gives a superficial overview, Goh (2009: 305), carried out a more localised and more detailed study into the attitudes of Chinese and Singaporean English language teachers to the question of teaching spoken language. Goh’s (2009) respondents were postgraduate students doing applied linguistic courses in the same university: 37 Chinese college and high school teachers and 38 (mainly) primary and secondary teachers from Singapore. The questions put to the respondents in an online forum were:

- Is linguistic information from British English data revealed in the CANCODE project useful for teaching learners about spoken language?
• Can knowledge of spoken grammar forms improve learners’ spoken language performance?

Goh (2009: 305) found interesting differences between the two groups:

They [Chinese teachers] consider the ability to speak naturally and accurately like a NS from countries such as Britain to be a distinct advantage. Among Singaporean teachers, however, opinion is divided, revealing the complexities in language choice in societies where the local English variety competes with Anglo models for allegiance and acceptance.

Finally, in terms of empirical research, Kuo (2006) reveals that many of her students see native speakers as a desirable model, particularly in relation to improving their chances in a globally competitive market.

The attitudinal research we have considered so far has largely been in relation to grammatical and lexical norms. There is also the question of which pragmatic and sociocultural norms should be conveyed through ELT materials and, as we shall see below, which norms of academic English should be applied by, for example, publishers and universities.

I would like to conclude this section with an anecdotal example of the practical issues raised by the increasing use of ELF. Some years ago I was involved in a European project developing an English language coursebook. I was the only native speaker on the project team which included members from six or seven different countries, but the meetings were always conducted in English. At one point, a discussion broke out among the team as to whether we needed an ‘agenda’ for the coursebook. I had no idea what they meant by ‘agenda’ in this context and it intrigued me that, as the only native speaker, I was the only person not able to understand the discussion. It turned out that they were referring to what I would call an ‘index’. This seemed, on the face of it, to be a good example of ELF in action and of the need for the minority native speaker to accommodate to ELF. On further reflection, however, I wondered how the discussion would have played out if the project team had included members from, for example, Japan, Brazil and India: they may have been more comfortable with the word ‘index’.

References


In the introduction to this book, I set out two main aims (italicised below):

Ideally, it [this book] would take us to a point where we stop talking about the potential of corpora to be applied to ELT, which has been proclaimed for many years, to a point where the ability to carry out corpus research and/or interpret the significance of corpus findings for the classroom is regarded as the norm for ELT practitioners.

In the course of this book, we have discussed the nature and pedagogic relevance of corpus findings in different domains: lexis, grammar, spoken grammar, ESP, learner corpora and ELF. In the domains of lexis, grammar and spoken grammar, while we considered a range of findings potentially relevant to ELT, we stressed that it is the overall picture of language which emerges from corpus research that we cannot ignore, a picture which shows the significance of lexical items such as collocations and chunks in communication and shows the need for a reappraisal of both the relationship between lexis and grammar and the relative importance of lexis and grammar in communication. The discussion of ESP and learner corpora took us closer to the classroom, while the discussion of teaching-oriented corpora and DDL took us into the classroom. The section on ELF underlined the need to assess the relevance of corpus findings in relation to corpus and context. I would like to claim, then, that the aims of facilitating corpus research and interpretation of corpus findings have been met. On reflection, I would add a third important aim: to enable teachers to support learners in using corpora autonomously – ultimately this may be the greatest contribution teachers can make to ensure a corpus-informed approach to ELT. My MA tutor once remarked to me, perhaps tired of my constant occupation of the middle ground, ‘Moderation in all things, including moderation’. I will be a little bolder, then, and argue that a critical awareness of what corpora have to offer in ELT is essential for any reflective ELT practitioner (I wish there was a function key ‘suppress hedging’, but I am doing my best here).

We also referred in the introduction to Römer’s (2006) observation that some ELT practitioners are resistant to corpus use in a pedagogical context. One reason
for this, I argue, is that it is a big step from being made aware of what corpora can do to ‘hands-on’ use of corpora. In this regard, I believe that teaching-oriented corpora such as SACODEYL, ELISA and BACKBONE have an important role to play in demystifying corpora: they have user-friendly interfaces and provide guidance for teachers as to how corpora can be exploited pedagogically. Resources such as these can perhaps revitalise the rather dusty image of DDL and, indeed, show that there is more to corpus-assisted language learning than concordance lines (though they certainly have their place).

We have considered another form of ‘resistance’ to corpora in ELT in the form of the charge that a corpus-informed approach to ELT unduly privileges the native speaker. As we have seen, however, learner corpora and ELF corpora are beginning to redress the balance. In any case, I would argue that this book promotes reference to corpora, not deference to corpora. Indeed, we have already discussed the notion of a ‘pedagogic filter’. A corpus is a reference resource; indeed, we might talk of a ‘corpus-referred’ approach (Timmis 2013) rather than a corpus-informed approach if we do not want to accord too much importance to the corpus.

The overall position we adopt in relation to the significance of corpora to ELT depends in large part, I would argue, on our answers to four questions posed by Ruehlemann (2008):

1. Do we have the time to teach how a feature is used differently in different context types?
2. Are we willing to rethink dearly-held convictions about the kind of English that we (should) teach?
3. Are we ready to take the trouble to familiarise ourselves with insights gained from corpus research?
4. What methodologies might be suited to help students come to terms with the inherent context-dependence of language use?

I hope this book will lead to ‘yes’ answers to the first three questions and a willingness to explore the fourth.

References


Appendices

Appendix 1: Commentaries

Commentary 1
1 Well
2 Think
3 Just
4 Right
5 Good
6 Money
7 House
8 Weather

you know (1)
I think (2)
a bit (3)
thank you (8)
in fact (11)
I suppose (15)
this morning (19)
very nice (84)

Commentary 2
Come up with collocates with mental products such as ideas, suggestions, conclusions etc.

Commentary 3
The node word is cause.
Commentary 4

They need to know that it is an irregular verb, an intransitive verb and its frequent collocates are, for example: *problems, difficulties, issues, questions* (but also, more positively, *opportunities*).

Commentary 5

D. uses rhetorical questions; hedges (a bit; I don’t really know); interactive discourse markers (you know); clause-final *though; I don’t really know*.

J: And the director of studies was like this dragon called xxxxx and she was sitting in the back of the room sort of observing me and I wasn’t following the strict regime so I got hauled into her office, you know and told that, ‘This is what you’ve got to teach, this is how we learn’ and you know really put you know really put to book whatever the expression is.

D: Are we any better today than that though? Aren’t we a bit like her?

J: Well, I just …

D: *Well*, ‘we’, I don’t really know, you know, isn’t it a bit the same, you know you come along you do your TEFL course and you’ve gotta do it this way this way.

D: *Well*, ‘we’, I don’t really know, you know, isn’t it a bit the same, you know you come along you do your TEFL course and you’ve gotta do it this way this way and then you go out and I’m sure directors of studies, cos there’s a bit of you know a power thing, isn’t there about all of that TEFL certification process that is self-perpetuating now because every there must be every director of studies in the world must have gone through a TEFL course by now and so what they look for is what they were told.

Commentary 6

Discourse marker *so*; ellipsis *went; discourse marker I think; hedging kind of; spoken pronunciation wanna, cos.

A: Yeah, so we got married by special licence on the Friday and went to Japan, I think, two weeks later.

B: Amazing.

A: I kind of realised that I didn’t really wanna teach kids ever, cos.

B: Very wise.
Commentary 7

Discourse marker *you know*; direct speech *come for an interview*; right you’ve got to go on a 15 hour training course.

J: Because I opened the paper and I saw an advert from Inlingua.
JB: Really!
J: …saying that they needed, you know, TEFL teachers, and er so I phoned them up and they said come for an interview so I flew to London cos in Jersey with mum and dad and they interviewed me and they said, ‘Right, you’ve got to go on a 15 hour training course’.

Commentary 8

I got a place at International House London, so we both did our course at the same time and then off I went so, *you know*, three week awful, awful, awful summer school in south of England, Hove, not Hove, somewhere awful.

D: It wasn’t EF was it?
H: No, it was even worse, I think and then I got, then *yeah* we decided, we had this grand master plan, me and Julie, we were gonna do a year in Paris, then we were gonna go for a year in Tokyo or Hong Kong, then we were gonna go for the third year in Rio de Janeiro and then we thought we’d *you know* think about settling down and being a bit more serious about life and so, *anyway*, that’s how I got into it. *So*, you *know*, we just packed a great big suitcase and went to erm went to Paris and *you know* paid a fiver a night for this, for these awful lodgings with, *you know*, broken panes in the windows and stuff like that and came down in our suits every morning.

Commentary 9

1 Comparator ‘like’
2 the verb ‘to like’
3 device to introduce reported thought ‘to be + like’
4 comparator ‘like’.

K: Well that’s the thing I mean when I was D.O.S-ing later in Dresden by then it wasn’t going to be my career it was another three years reflecting on life I found out I was *like* (1) in the hierarchy if you *like* (2) and it began to annoy me when people would turn literally kind of walk through the door and say Oh you know I’m passing through can I teach here for a few weeks Dresden/Prague/Vienna And I was *like* (3) well NO you know this is a career I would hope you know this is something it’s *like* (4) these people were just
passing through but THEN that’s the problem you see we’re all passing through we’re here for six weeks six months six years ( )

**Commentary 10**

**Conversation 1**

S1: How do I get to the Vet Clinic?
S2: Oh, my God. It’s far, but, um. It’s on, it’s Shaw Lane is right, if you keep going up here you’ll hit Shaw Lane and it’s going that way. But go a little bit more and then cross the parking lot and it goes that way.
S1: Umhum.
S2: And then cross the road and just follow that sidewalk all the way down. If you keep following it you’ll see a sign that says Vet Clinic. But go all the way down there and cross the road and just follow the sidewalk down.
S1: Okay.
S2: Okay?
S1: Thanks a lot

**Conversation 2**

S1: How do I get to the Vet Clinic?
S2: **The Vet Clinic?** Oh man, it’s a haul (pause) Uh, okay. Go up to the bridge, and that’s on Farm Lane. And you want to go right all the way up until you pass—um. Do you know where—um. Are you familiar with the campus?
S1: Pretty much.
S2: Do you know where the (pause) Uh Agriculture Hall is?
S1: Yeah, I know.
S2: Okay. When you see the Agriculture Hall, it will be on your right hand side. You want to go left again down towards Akers. Then the Vet Clinic should be on that road, and it says Vet Clinic.
S1: **It does?**
S2: Yeah. You can’t miss it. It should be just past the intersection of Bogue Street and the street you’ll be on (rising intonation).
S1: Okay.
S2: All right.
S1: Thank you very much.

**Commentary 11**

- Ellipsis (numerous examples)
- Direct speech
- Discourse markers ‘like’, ‘I mean’, ‘you know’
- Vague language ‘*this, that and the other*’
- Discourse marker ‘I mean’
• Hedging ‘sort of’
• Question tag ‘weren’t they?’
• Clause-final use of though

J: Yeah, so then I just looked in the newspaper, thought, ‘Well shall I just go into primary school teaching or…’ then I saw this job interview in Portugal so, just decided to go to the interview and went and ended up going down to this hotel in London into a bedroom being interviewed and no qualifications and nothing at all and they offered me like free Portuguese lessons and free flight and free this that and the other, free accommodation so I just thought, ‘Well is it primary school teaching or is it jump on a plane and head off to Lisbon so that’s why I did it and also I mean didn’t ever think I’d sort of end up with it, you know, as a career or anything at all
I: But they were quite a decent school to work for though weren’t they, I’m surprised.

Appendix 2: Corpora and corpus resources

Large Open Access Corpora
British National Corpus: http://corpus.byu.edu/bnc/
Corpus of Contemporary American English: http://corpus.byu.edu/coca

Spoken Corpora
Corpus of American Soap Operas: http://corpus.byu.edu/soap/
Available at: http://www.linguistics.ucsb.edu/research/santa-barbara-corpus#Intro

EAP Corpora
British Academic Written English Corpus: http://www.coventry.ac.uk/research/research-directory/art-design/british-academic-written-english-corpus-bawe
British Academic Spoken English Corpus: http://www2.warwick.ac.uk/fac/social/research/collect/base/
Michigan Corpus of Academic Spoken English: http://quod.lib.umich.edu/m/micase/

Learner Corpora
English Profile Corpus: http://www.englishprofile.org/index.php/corpus
**ELF Corpora**

Vienna and Oxford International Corpus (VOICE): http://www.univie.ac.at/voice/

English as a Foreign Language in Academic Setting (ELFA): www.helsinki.fi/elfa/elfacorpus

**Pedagogic Corpora**

System Aided Compilation and Distribution of European Youth Language: http://sacodeyl.inf.um.es/sacodeyl-search2/

Backbone – Corpora for Content & Language Integrated Learning http://projects.ael.uni-tuebingen.de/backbone/moodle/

English Language Interview Corpus a a Second-Language Application: http://www.uni-tuebingen.de/elisa/html/elisa_index.html

**Corpus tools and resources**

Antconc concordance: http://www.laurenceanthony.net/software.html

Wordsmith tools: http://www.lexically.net/wordsmith/

Academic Word Highlighter: http://www.nottingham.ac.uk/alzsh3/acvocab/awlhighlighter.htm

Lexical Tutor: http://www.lextutor.ca/

Corpus Tagger: http://ucrel.lancs.ac.uk/claws/trial.html
Index

Academic Collocation List (ACL) 51, 154, 156, 172–3
Academic Corpus 151
Academic Formulas List (AFL) 51, 155, 172–3
Academic Vocabulary List (AVL) 151–2
Academic Word Highlighter 174, 203
Academic Word List (AWL) 44–5, 137, 151–3, 162–3, 172, 177
Ackermann, K. 47, 154–5, 173–4
activity types 12
ad hoc corpora 10, 147–8, 162
Adel, A. 177
adjectival pre-modification 59
Aktas, R. 153
Allan, R. 132, 136
alphabetical lists 17, 19
Alptekin, C. 191
annotated corpora 121, 129–30
Antconc concordance 17, 27, 132, 203
applied linguistics 4
approximation 89
Asian Corpus of English (ACE) 186, 189
Aston, G. 61, 136, 140, 142, 184
authenticity 113–14, 133, 135, 138–41, 146, 176, 193
autonomy 135, 141–2, 147–8, 174–5
awareness-raising 67, 70, 76, 111, 174
back-channelling 113–14
BACKBONE 130, 197, 203
backshift rules 60, 62–3
Bank of English Corpus 9
Barbieri, F. 57, 62–4, 67, 70
Bennett, G. 121
Bernadini, S. 139
Bernsten, J. 100
Biber, D. 26, 57–61, 71, 74–5, 84, 87, 156
Bielsa, S. 125
bilingualism 122, 191
Biskup, - 47
Bjorkman, B. 187
Bolton Corpus 91, 93, 96
Boulton, A. 136, 139, 141
Boxer, D. 102–3
Braun, S. 128–9
Brazil, D. 2
British Academic Spoken English (BASE) 149–50, 180, 202
British Academic Written English (BAWE) Corpus 136, 148, 150, 179, 202
British National Corpus (BNC) 2–3, 10, 17, 38–40, 43; classroom approaches 119, 132–3, 139, 141; and ESP 148, 155, 163, 167; and grammar 64, 66, 68, 78; and lexis 45, 49–50, 53; resources 202; spoken corpus research 82
Brown Corpus 45, 64
bundles 24, 26–7, 34, 45, 50, 156–9, 172
Burton, G. 9, 128
Business English 147–8, 166–71, 186
Buttery, P. 121, 126, 128

CALL programmes 128
Cambridge English Profile Corpus (CEPC) 15, 122–3, 125
Cambridge International Corpus (CIC) 9, 66, 169
Cambridge Learner Corpus (CLC) 9, 121, 127
Cambridge and Nottingham Business English Corpus (CANBEC) 3–4, 169
Cambridge and Nottingham Corpus of Discourse English (CANCODE) 3, 83, 88, 169–70, 193
Camiciottoli, C. 176
Capel, A. 127
Cardinal Concessive Schema 100
Carter, R. 2, 8, 15, 19, 36, 41–3, 52, 57, 66, 72, 84–6, 88, 90, 93–7, 106–7, 110, 112, 115, 169–70, 192
case studies 58–69, 91, 120, 162, 166, 174
Chambers, A. 140
Channell, J. 90
Check My Words 138
Chen, Y.-H. 47, 154–5, 173–4
Cheng, W. 164, 166
Cheshire, J. 92
Chomsky, N. 183
chunks 1, 23–4, 26–7, 29–32, 34–7; classroom approaches 132; and ESP 150, 154–5, 157, 169–70, 172, 174; and grammar 76; and lexis 39, 41–2, 45, 49–54; perspectives 188; role 196; spoken corpus research 88, 90, 106–7, 109–10
classroom practice 119–45, 189–94
CLT 81, 135
Cobb, - 47
CoBuild 9, 65
codes 16–17, 167, 188
Cogo, A. 161, 167, 185–8
colligation 24
colloctions 1, 4–5, 7–8, 17, 19–21; classroom approaches 134, 136–8; commentaries 198–9; definition 24; and ESP 150, 154–5, 169–70, 173–4, 180; and grammar 74; and lexis 23, 26, 29–32, 34–42, 45–51, 53–4; perspectives 185, 188; role 196; spoken corpus research 88–90, 104
commentaries 198–202
Common European Framework (CEFR) 123, 126–8, 130, 132
Comparison Corpora 155
competence 183
concgramOnline 165
concgrams 27–8, 165–6
concordances 9–10, 17–19, 27, 36, 39; classroom approaches 122, 129, 133, 135–41; and ESP 165, 174–6, 179–80; role 197
conditionals 62, 64–7, 77
Coniam, D. 174
connections 71–7
Conrad, S. 1, 29, 31–2, 61, 68, 70–1, 74–5, 77, 84, 98
consciousness-raising (CR) 133, 173–4, 192
Constituent Likelihood Automatic Word-Tagging System (CLAWS) 17, 121
Content and Language Integrated Learning (CLIL) 10
context 3, 5–6, 8, 12, 15; classroom approaches 120, 122, 126, 131, 135–6, 138–9, 142; and ESP 148, 155, 161–2, 167, 169, 172, 174–5, 177–9; and grammar 52, 63, 76; and lexis 23, 30–1, 36, 41, 44; perspectives 184–6, 188–9, 193–4;
<table>
<thead>
<tr>
<th>Term</th>
<th>Page Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>role</td>
<td>196–7; spoken corpus research 81–3, 90, 97, 103, 108, 110, 112–13</td>
</tr>
<tr>
<td>conversation</td>
<td>6, 15–16, 42–3, 59–60, 201; classroom approaches 123–4; and grammar 63, 68, 75; research 83–4, 86–7, 90, 92, 95–6, 98–100, 104, 107, 111, 114</td>
</tr>
<tr>
<td>Conzett, J.</td>
<td>32, 47–8</td>
</tr>
<tr>
<td>Cook, G.</td>
<td>7, 114, 184, 188</td>
</tr>
<tr>
<td>corpus 1; building 14–21; data analysis</td>
<td>17–21; definitions 2–3; discussions 12, 14, 27, 34, 40, 43, 45, 47, 51, 54, 57, 59, 62, 64, 67, 74, 81, 84, 86, 88, 91, 99, 103, 107–8, 110, 112, 116, 120–1, 124, 128, 130, 132, 142, 147, 150, 153, 155, 157–9, 161, 164, 166–7, 170–1, 184–8, 192, 194, 196; and ELT 7–12; limitations 183–4; perspectives 34–53, 183–95; qualitative analysis 6–7, 23, 74, 85, 184, 187; quantitative analysis 4–6, 23, 41, 85, 187; questions 4–6, 9, 12, 34, 37–8, 40, 53, 85–7, 89–90, 93, 100, 115; quick-and-dirty 10, 12, 120, 131–2, 147, 176; role 196–7; searches 4, 10, 12, 16–17, 21, 24, 30, 32, 34, 36, 43–4, 78, 92, 130, 136–7, 141, 150, 164, 168, 179; types 3–4</td>
</tr>
<tr>
<td>Corpus of American Soap Operas</td>
<td>84, 202</td>
</tr>
<tr>
<td>Corpus of Contemporary American English (COCA) 2–3, 17–18, 63, 82, 133, 139, 151–2, 202</td>
<td></td>
</tr>
<tr>
<td>Corpus Encoding Standard (CES) 16 corpus structures see structures</td>
<td></td>
</tr>
<tr>
<td>Corpus Tagger 203</td>
<td></td>
</tr>
<tr>
<td>corpus-aided language learning</td>
<td>142</td>
</tr>
<tr>
<td>Cortes, V.</td>
<td>153</td>
</tr>
<tr>
<td>Couper-Kuhlen, E.</td>
<td>100</td>
</tr>
<tr>
<td>coursebooks 3, 9, 31, 51, 62; classroom approaches 128, 131; and ESP 177; and grammar 70, 75; perspectives 191, 194; spoken corpus research 91, 101–3, 106–7, 112–13</td>
<td></td>
</tr>
<tr>
<td>Coxehead, A.</td>
<td>151–3, 172, 177</td>
</tr>
<tr>
<td>Crawford, W.</td>
<td>92</td>
</tr>
<tr>
<td>criterial features</td>
<td>123, 127</td>
</tr>
<tr>
<td>cross-cultural comparisons</td>
<td>52</td>
</tr>
<tr>
<td>Cullen, R.</td>
<td>112</td>
</tr>
<tr>
<td>data-driven learning (DDL)</td>
<td>10, 14, 119, 132–42, 146, 174–7, 179, 196</td>
</tr>
<tr>
<td>databases</td>
<td>16, 119, 133, 141, 184</td>
</tr>
<tr>
<td>Davies, M.</td>
<td>151–2</td>
</tr>
<tr>
<td>De Cock, S.</td>
<td>15, 120–1, 126</td>
</tr>
<tr>
<td>De Haan, P.</td>
<td>125</td>
</tr>
<tr>
<td>DeCarrico, J.</td>
<td>34–5</td>
</tr>
<tr>
<td>delexical verbs</td>
<td>30</td>
</tr>
<tr>
<td>Dellar, H.</td>
<td>114</td>
</tr>
<tr>
<td>demography</td>
<td>3, 5, 15, 82, 123, 183</td>
</tr>
<tr>
<td>descriptions</td>
<td>57–8, 62, 64, 66, 70–1; classroom approaches 119–25, 129, 135–6; and ESP 148, 151, 159, 164, 167; and grammar 77; perspectives 186, 190; spoken corpus research 82–4, 115</td>
</tr>
<tr>
<td>design criteria</td>
<td>120, 129</td>
</tr>
<tr>
<td>developmental sequence</td>
<td>126–7</td>
</tr>
<tr>
<td>Dewey, M.</td>
<td>161, 167, 185–8</td>
</tr>
<tr>
<td>dialect</td>
<td>81, 114, 192</td>
</tr>
<tr>
<td>dictionaries</td>
<td>9, 120, 163, 177</td>
</tr>
<tr>
<td>discourse</td>
<td>52, 64, 68, 70, 81; classroom approaches 123, 126, 129; commentaries 199–201; and ESP 155–6, 169, 178; spoken corpus research 85–9, 93–4, 96–102, 104, 106, 108, 115</td>
</tr>
<tr>
<td>Durham, M.</td>
<td>96</td>
</tr>
<tr>
<td>Durrant, P.</td>
<td>22, 43, 154</td>
</tr>
<tr>
<td>Eckhardt, S.</td>
<td>57, 62–4, 67, 70</td>
</tr>
<tr>
<td>EFL</td>
<td>47, 49, 63–4, 125, 131</td>
</tr>
<tr>
<td>electronic texts</td>
<td>2</td>
</tr>
<tr>
<td>elicitation</td>
<td>77</td>
</tr>
<tr>
<td>ellipsis</td>
<td>94–5, 106–7, 183, 199, 201</td>
</tr>
</tbody>
</table>
Ellis, N. 51, 155–6, 173
ELT 1, 3, 8, 16, 196–7; classroom approaches 136, 142; and ESP 146, 169; and grammar 58–60, 62–5, 67–8, 70, 75; and lexis 22–3, 29, 31–2, 34, 42–3, 51, 53–4; perspectives 183–4, 188, 190–1, 193; spoken corpus research 81–2, 85, 91, 97, 100, 102–3, 115
engagement features 159
Engineering English 147–8, 161–6
Engineering Technology Word List 163
English for Academic Purposes (EAP) 126, 142, 146–8, 151–61, 174–8, 180, 186, 202
English for General Academic Purposes (EGAP) 146
English Language Interview Corpus as a Second Language (ELISA) 129, 197, 203
English as a Lingua Franca in Academic Settings (ELFA) 149, 161, 203
English as a Lingua Franca (ELF) 130, 161, 167, 171, 183–90, 194, 196, 203
English Profile Corpus 128, 202
English for Specific Academic Purposes (ESAP) 146
English for Specific Purposes (ESP) 3, 10, 81, 122, 146–82, 196
ENL 187–8
error tagging 121–2
ESL 49, 63–4, 70, 185
evaluation 51, 68, 87, 96, 99, 107, 156, 159, 171, 180
Evison, J. 99
examinations 9, 15
existential there 91–6
expert writing corpora 147–8, 171
face2face 9
Farr, F. 10
fitness for purpose 2
FLOB 155
fluency 34–6, 41, 76, 106–7, 126, 155, 188
formulaic sequences 22, 24, 26, 49, 155
framing signals 159
Francis, G. 71–2, 75–6
Frazier, S. 64–5
frequency 4–5, 9, 16–17, 21, 41–52; classroom approaches 128, 132, 136, 138, 141; and ESP 151–4, 157–8, 162, 167, 169–70, 172–4; and grammar 58–62, 67, 70, 77; and lexis 54; spoken corpus research 84–91, 104
Frown Corpus 155
gap-fills 44, 135, 137
Gardner, D. 151–2
Gavioli, L. 61, 184
GCSEs 10
General Service List (GSL) 152, 162–3
genre-specific corpora 82, 140
genres 3, 5, 14–15, 48, 75; classroom approaches 120, 131–2, 140; and ESP 150–1, 156, 159–60, 169; perspectives 183; spoken corpus research 82, 84
Gilmore, A. 113
Gilquin, G. 121, 140–1, 147, 151, 177
Goh, C. 193–4
Götz, S. 141
Gouverneur, C. 51
Gralla, L. 125
grammar 7, 9–10, 17, 22–3, 27; classroom approaches 121, 123, 125–9, 131, 133, 136, 138, 141; and ESP 152, 154, 159, 166, 171, 175–6, 178; and lexis 29, 31–2, 34–5, 43, 46, 54; perspectives 185,
Index

189–92, 194; and research 57–80; role 196; spoken corpus research 81–2, 91–8, 106–8, 112, 115
Grammar McNuggets 61, 102
Granger, S. 119–22, 126–8, 140–1, 147, 151, 177
Greaves, C. 22, 27–8
Green, A. 123

Hadley, G. 133
Handford, M. 170
Haswell, R. 154
Hawkins, J. 121, 126, 128
head structure 52, 106–7, 112, 178
Hedevind, B. 93
hedging 52, 88–9, 106, 126, 156, 186, 196, 199, 202
Hewings, M. 138, 147, 171, 176
highlighters 44–5, 51, 137, 174
Hill, J. 48
Hirvela, A. 174
Hoey, M. 41
Holmes, J. 101–2
Hong Kong Engineering Corpus (HKEC) 3, 162, 164–5
Hong Kong Financial Services Corpus 166
Hong Kong Institution of Engineers (HKIE) 164
Hsieh, W.-M. 160
Hsu, J. 51
Huang, L.-S. 142
Hughes, R. 57
Hunston, S. 71–2, 75–6, 183–4
Hyland, K. 126, 152, 155, 157–8

ICE 70, 88
idioms 34, 52, 99, 154, 183, 187–8
if-clauses 64–5, 67
Illustration-Induction-Interaction (III) 110
improvised corpora 131–2
India 3

induction 110
industry standards 17
INTELeNG Corpus 125
interaction types 83–4, 86–8, 99, 110, 139, 178
International Corpus of Learner English (ICLE) 3–4, 121–2, 124–5, 202
International English 190–1
internet see websites
interpretations 184
intertextual coherence 129
intuition 22, 24, 30–1, 41, 156, 175–6
Japanese EFL (JEFL) Corpus 120
Japanese language 120
Japanese Learner English (JLE) Corpus 122–5, 202
Jarvis, - 174
Jenkins, J. 161, 167, 185–8, 190
Johns, T. 10, 119, 135–6
Jones, C. 64–7, 70, 111
Jones, M. 22, 43
Journal Paper Corpus 160

Kachru, - 193
Kirkpatrick, A. 186
Koprowski, M. 51
Krishnamurthy, R. 22
Kuo, I-Chun 112, 194

L1 language development 10, 120–6, 128, 153, 189, 191
L2 language development 49, 129, 153, 177
Lambrecht, K. 96
Larsen-Freeman, D. 103
Latin language 92
Lau, H. 159–60
Lawson, A. 57
layers of meaning 36–41
learner corpora 3, 10, 119–26, 128, 133; classroom approaches 125–8,
Index

140; and ESP 147–8, 171–2, 177, 180; perspectives 184; resources 202; role 196
learners 14–15, 36, 41–2, 45, 47–9; classroom approaches 119–21, 123–6, 128–9, 131–6, 138–42; and ESP 150, 153, 159, 171–2; and grammar 59–61, 67, 74–7; perspectives 184–6, 190–3; spoken corpus research 97, 102–4, 106, 108, 110–12, 115
learning theory 138
Lee, D. 147, 174–5
Leech, G. 3, 8, 46, 119
lemmas 24, 44–5, 152, 162
Lewis, M. 23, 47, 52, 90
lexical bundles see bundles
lexical chunks see chunks
lexical priming see priming
Lexical Tutor 203
lexis 7, 9, 187, 189–90, 194;
classroom approaches 123, 126, 128–9, 132–3, 136, 138; and ESP 150–9, 162–3, 166, 168–79; and grammar 57–8, 61, 71–7; role 22–56, 196; spoken corpus research 81–2, 84–106, 108, 114
like 63–4, 75, 85, 93–4
Limerick Corpus of Irish English (L-CIE) 82–3
lingua franca 3, 49, 130, 149, 161, 167, 183–91, 194, 196
Lingua Franca Core (LFC) 190–1
Liou, H.-S. 160
location 158
LOCNEC 124
logistics 138, 176
Longman Corpus of Spoken American 83
Longman Spoken and Written English (LSWE) 62–3
Louvain International Database of Spoken English Interlanguage (LINDSEI) 122, 124
McCarten, J. 9, 43, 70, 88, 111, 113–14
McCarthy, M. 2, 8–9, 15, 19, 36, 41–3, 52, 57, 66, 70, 72, 84–6, 88, 90, 93–9, 106–7, 110–15, 169–70
McEnery, T. 2, 16
McGraw-Hill 163
Manning, E. 72, 75–6
mark-up schemes 16
Martinez, R. 49–51, 155
materials 14, 49, 62, 65, 67; classroom approaches 119, 125, 130–1, 134–6, 140; and ESP 146, 148, 151–2, 162, 169, 172, 176, 178–80; and grammar 75; perspectives 189; spoken corpus research 102–10, 113–15
Maule, D. 65
Mauranen, A. 111
Mendikoetxea, A. 125
metadata 16
metadiscourse 156, 171
metaphor 42, 53, 61, 72, 104
methodology 4, 9, 70, 81–2, 85, 103–15, 148, 156, 173, 192, 197
metonymy 42
Meunier, F. 128, 140
Michigan Corpus of Academic Spoken English (MICASE) 3, 64, 139, 149, 155, 178, 180, 202
micro-corpora 131
MICUSP Corpus 177
Miller, J. 93–4
Milton, J. 126, 128
Mindt, D. 62, 70
Modiano, M. 191
Monoconce Pro 14
monolingualism 119–20, 122–3
Moon, R. 36–7, 42, 51–2
Mudraya, O. 179
Mukherjee, J. 140
multilingualism 120–2
Mumford, S. 107
music teaching English 10–12, 14
narrative tenses 97
Nation, P. 26, 42, 45–6, 48, 50, 88
native speakers 35–6, 41, 47, 81, 197;
classroom approaches 121–2, 124–6, 128–9, 140; and ESP
148–50, 172, 175, 177; perspectives 183–94; spoken corpus research
102–3, 107
Nattinger, J. 34–5
Nelson, M. 167–8, 170–1
Nesi, H. 147
Nesselhauf, N. 47
new quotative devices 63–4
Ng, Y.-J. 163
gngrams 27, 33, 132
NICT 122–3, 202
node words 30, 38, 45, 48, 198
noun pre-modification 59, 61

Objective First Certificate 9
O’Keeffe, A. 2, 8, 10, 15, 19, 27, 36, 41–3, 52, 57, 72, 84–6, 88, 90, 99, 106, 139, 169–70
open choice principle 34
organisational frameworks 27

Paquot, M. 121, 147, 151, 177
paralinguistics 184
participant-oriented features 159
Partington, A. 36–8
parts of speech (PoS) 17, 25, 30, 72, 129
pattern grammar 75–7
pausing devices 107
Pearson International Corpus of
Academic English (PICAE) 150, 154
pedagogic corpora 3, 67, 131, 147–8, 161–3, 174, 203
pedagogy 1–3, 8, 16, 24, 27–8;
classroom approaches 119, 123, 128–30, 135, 140; and ESP 153, 155, 157–8, 160–1, 163, 173, 177; and grammar 58, 60–2, 64–7, 72, 75, 77; and lexis 32, 34, 41, 43,
45–6, 48–9, 51–2, 54, 58;
perspectives 190; role 196–7;
spoken corpus research 83, 103, 106, 115
perspectives 57–8, 62–4, 183–95
Peters, A. 35
phrasal expressions 49–50
Pickering, L. 102–3
Pitzl, M.-L. 187–8
pivot words 45
politeness 89, 101–2, 106, 114
polysemy 44
Pomerantz, A. 99
pragmatics 4, 6–7, 26–7, 52–3, 64;
classroom approaches 129, 139; and ESP 156, 172, 178; perspectives 189, 194; spoken corpus research 86, 89–90, 96, 101, 103, 106
Pravec, N. 120–1
pre-modification 59, 61
precision 140
prefabricated units 24, 35, 92
preference 23, 36, 38, 40–1, 54, 99, 134, 191
prescriptions 62
present-practise-produce (PPP) 61, 70, 102, 111
preservation of face 89–90, 106, 114
priming 36, 41
procedure 159
Prodromou, L. 185
progressive verb forms 59–61, 71, 75
pronunciation 176
prosody 36, 38–41, 54, 72, 74, 102, 170, 184
proverbs 52
puns 114
qualitative analysis 6–7, 23, 74, 85, 184, 187
quantitative analysis 4–6, 23, 41, 85, 187
quotative devices 63–4
rank order lists 17
Rankin, I. 97
Ranta, E. 161
Rayson, P. 46
recall 140
reference books 9, 128
referential expressions 156
reflex fallacy 103
relevance of findings 81
reported speech 58, 60, 62–3, 67, 93, 97, 107, 185
Reppen, R. 57–60, 75
research 1–2, 9, 16, 22–4, 26–9;
classroom approaches 119, 121–3, 125–9, 131, 136, 141; ELF 185–9;
and ESP 146–7, 151–61, 164, 171, 177; and grammar 57–80; and lexis 31–2, 34, 36, 47, 54; perspectives 184, 191, 193–4; role 196–7;
spoken language 81–118
Reshöft, N. 125
resources 14, 120, 138, 140, 146, 176–7, 197, 202–3
response tokens 86–7, 99, 113
resultative signals 159
right dislocation 95
Rohrbach, J. 140
Rollinson, P. 125
Römer, U. 1, 3, 8, 23, 29, 31–2, 60, 62, 70–1, 131, 196
Ruehlemann, C. 39, 70, 92, 197
Sandiford, H. 70
Santa Barbara Corpus of Spoken American English 64, 83
saving documents 2, 29, 31
Schmitt, N. 26, 49–51, 155
Scott, M. 176
Scotton, C. 100
second language acquisition (SLA) 4
Seidlhofer, B. 188
selection criteria 52, 61–2, 102–8, 112, 120, 139, 155–6, 162–3, 187
semantics 21, 23, 26–7, 36–41, 49–50;
classroom approaches 123, 134, 139; and ESP 167–8, 170, 172; and grammar 72, 74–5, 78; and lexis 52, 54; perspectives 187; spoken corpus research 106
sequence 108–10, 156, 160
Shin, D. 26, 45–6, 50, 88
Shortall, M. 70
simple verb forms 59–61, 75
Simpson-Vlach, R. 51, 155–6, 173
Sinclair, J. 7, 16, 22, 34, 138
slang 81
slot and filler approach 34
Soars, J. 67
Soars, L. 67
Sobin, N. 92
sociocultural norms 83, 104, 108, 115, 184, 190, 194
software 2, 14, 16–17, 27, 29, 31, 33, 121, 133, 136, 138, 165
Spanish language 10
specialisms 10, 146
spoken language 2–3, 14–15, 19, 45, 53; classroom approaches 120, 123, 125–6, 139; commentaries 199; and ESP 147–9, 155, 167, 175; and grammar 57–8, 62–4; perspectives 184, 193–4; research 81–118;
resources 202; role 196
SST 124
stance 156, 159
storage 16
Stranks, J. 63, 75
structures 1, 4, 22–3, 35, 37, 40–1, 51–2, 54, 57–8
structuring signals 159
Stubbs, M. 71
Swales, J. 147, 172–5
Switchboard Corpus 84, 155
syllabus design 3, 9, 14, 42–4, 46–7;
classroom approaches 120, 128, 131; and ESP 148, 172–3, 176, 180; and grammar 58, 61–2, 77; and
lexis 51–3; perspectives 191–2; spoken corpus research 91, 97, 102–10, 115
syntax 71, 102, 106, 121, 123
System Aided Compilation and Distribution of European Youth Language (SACODEYL) 3–4, 130, 140, 197, 203
tagging 17, 121–2, 129
Tagliamonte, S. 92
tail structure 58, 95–6, 106–7
Tampere University of Technology 149
Tao, H. 98
TBL 81
teacher training courses 61, 67
Teachers Telling Tales (TTT) 15, 85–6, 89, 93, 98, 108
teaching-oriented corpora 3, 119, 128–33, 140
TEFL 85, 87
terminology 24–34
TESOL 180
Text Encoding Initiative (TEI) 16
textbooks 49, 60, 63–4, 70, 113, 128, 131, 148, 154, 162–3
texts 2–3, 5, 14–16, 29–30, 41–2; classroom approaches 119, 121, 129, 131–3, 136–40; and ESP 151, 158–9, 162, 166, 174–5, 177, 180; and grammar 77; and lexis 44; spoken corpus research 83, 96, 103, 108, 110, 112
Thompson, S. 100, 172, 180
Thornbury, S. 61, 102
though 62, 68–70, 98–9, 199, 202
threshold targets 42
Timmis, I. 6, 54, 57, 103–4, 106, 108, 110–11, 113, 184, 193
TOEFL 63, 174
Tomlinson, B. 108, 152
Tono, Y. 16, 120
topics 159
Touchstone 9, 70, 111
transcription 10, 15, 82, 123–4, 139, 149, 154, 178
transition signals 159
Tribble, C. 10, 131
Tse, P. 152
turn-taking 92, 113–14
tutorials 2, 29, 31, 33
University of Louvain 122
University of Michigan 149
University of Tampere 149
Ur, P. 66
vagueness 89–91, 106, 126, 156, 201
Van der Haagen, M. 125
variable concord 91–6
Vienna-Oxford International Corpus of English (VOICE) 3–4, 185–6, 189, 203
vocabulary 1, 23, 29, 31–2, 35; classroom approaches 127–8, 131, 136, 138; and ESP 151–2, 162–3, 169–70, 172, 176, 178; and grammar 71, 77; and lexis 42–4, 46, 49, 52–3; spoken corpus research 84, 86, 106, 109
Waller, D. 64–7, 70
Walsh, S. 113, 136
Ward, J. 147, 162–3, 173
Warren, M. 22, 27–8, 164, 166–7
Weinert, R. 93–4
Widdowson, H. 61, 139–40, 191
Wilkins, D. 23
<table>
<thead>
<tr>
<th>Name</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Willis, D.</td>
<td>3, 35, 65, 70, 108, 110, 131, 185, 190–2</td>
</tr>
<tr>
<td>Wilson, A.</td>
<td>2, 46</td>
</tr>
<tr>
<td>wordandphrase.info</td>
<td>174</td>
</tr>
<tr>
<td>WordPilot programmes</td>
<td>128</td>
</tr>
<tr>
<td>words</td>
<td>22–7, 30, 32, 34, 38; classroom approaches 127–8, 132, 137–8; and ESP 151–3, 162–4, 167–9, 172; and grammar 61, 72, 74, 76–7; and lexis 41–6, 48, 50, 54; spoken corpus research 84, 88, 104</td>
</tr>
<tr>
<td>Wordsmith Tools</td>
<td>14, 17, 163, 203</td>
</tr>
<tr>
<td>written language</td>
<td>2–3, 14–15, 57, 59–60, 62–3; classroom approaches 120, 125–6, 128, 136, 138; and ESP 147–8, 151, 153, 158–60, 167, 172, 174–5, 177; spoken corpus research 82, 84, 86, 88, 91, 93, 95–6, 98, 115</td>
</tr>
<tr>
<td>Xiao, R.</td>
<td>16</td>
</tr>
<tr>
<td>Yoon, H.</td>
<td>174</td>
</tr>
</tbody>
</table>
eBooks
from Taylor & Francis
Helping you to choose the right eBooks for your Library

Add to your library’s digital collection today with Taylor & Francis eBooks. We have over 50,000 eBooks in the Humanities, Social Sciences, Behavioural Sciences, Built Environment and Law, from leading imprints, including Routledge, Focal Press and Psychology Press.

Choose from a range of subject packages or create your own!

Benefits for you
- Free MARC records
- COUNTER-compliant usage statistics
- Flexible purchase and pricing options
- 70% approx of our eBooks are now DRM-free.

Benefits for your user
- Off-site, anytime access via Athens or referring URL
- Print or copy pages or chapters
- Full content search
- Bookmark, highlight and annotate text
- Access to thousands of pages of quality research at the click of a button.

Free Trials Available
We offer free trials to qualifying academic, corporate and government customers.

eCollections
Choose from 20 different subject eCollections, including:

- Asian Studies
- Economics
- Health Studies
- Law
- Middle East Studies

eFocus
We have 16 cutting-edge interdisciplinary collections, including:

- Development Studies
- The Environment
- Islam
- Korea
- Urban Studies

For more information, pricing enquiries or to order a free trial, please contact your local sales team:

UK/Rest of World: online.sales@tandf.co.uk
USA/Canada/Latin America: e-reference@taylorand francis.com
East/Southeast Asia: martin.jack@tandf.com.sg
India: journalsales@tandf india.com

www.tandfebooks.com