Corpora in Language Teaching

Chapter · November 2009

DOI: 10.1002/9781444315783.ch19

1 author:

John Flowerdew
Lancaster University
174 PUBLICATIONS **5,580** CITATIONS

SEE PROFILE

19 Corpora in Language Teaching

JOHN FLOWERDEW

Introduction

Applications of corpus linguistics to language teaching began in the late eighties and early nineties. Examples of early work are Higgins and Johns (1984), Higgins (1988), Johns (1988, 1991), Tribble and Jones (1990), Stevens (1991), and J. Flowerdew (1993a). Most work in this area, as in other areas of applied linguistics and language teaching, has focused on English. However, some examples on other languages are Wichmann (1995), Ahmed and Davies (1997), Dodd (1997), King (1997), Kennedy and Miceli (2001), Rule et al. (2003), Belz (2004), Rule (2004), and Bolly (2005). Previous overviews of the field are Leech (1997), Aston (2001), Biber and Conrad (2001), Bernardini (2004), and Stubbs (2004).

Interest in corpus-based approaches to language teaching has developed quite rapidly in recent years, so that now there is a wealth of literature and, although less, still considerable application in this area. Application at the level of primary and secondary schools, however, has not kept pace with the considerable developments that are now going on at the tertiary level, especially in languages for specific purposes (LSP). Gavioli (2005), for example, is only able to cite one project at primary level (Sealey & Thompson, 2004).

One of the reasons for the relatively slow rate of classroom application has been the limitations of the technology. However, as Leech stated already in 1997, "computers have grown smaller, cheaper, and massively more powerful" (1997, p. 2). Since that statement, this trend has continued. In addition, more and more corpora have become available and it is easier to create personalized corpora. Furthermore, people are becoming increasingly computer literate and are therefore more easily introduced to the new approach. But if the start has been rather slow, this has a positive side, in that corpus applications to pedagogy have avoided, to quote Leech (1997, p. 4) again, the "bandwagon" effect. In developing more slowly, the risk of corpus-based approaches to language teaching following the path of the language laboratory, for example, with its its meteoric rise and ultimate demise, may be avoided.

The Handbook of Language Teaching Edited by Michael H. Long and Catherine J. Doughty
 ISBN: 978-1-405-15489-5

What are the Principles in Corpus Linguistics that Can Be Applied to Language Teaching?

A corpus is a large database of language. Although the first corpora were relatively small – the Brown corpus (developed at Brown University, USA in the early 1960s) consisted of one million words – there now exist corpora consisting of hundreds of millions of words (e.g., the British National Corpus (BNC), 100 million words; and the Bank of English (COBUILD at Birmingham University, UK), over 500 million words). At the same time, however, much smaller corpora with as few as 100,000 words or less are being created all the time for specialist applications. It should be borne in mind, however, that, as pointed out by Gavioli and Aston (2001, p. 238), even the very large corpora contain less language than the average user will have experienced in their daily life.[1] In addition, the linguistic content of corpora is different from what is experienced by individuals in real life, many of them consisting largely of written language. Furthermore, while each text is given equal weighting in a corpus, in real life some texts will hold more value and be experienced more times than others (poetry and religious texts, for example, might be highly valued and heard or read many times). While some corpora are kept in a "raw" state (e.g., Bank of English), many are "tagged" (i.e., coded, according to parts of speech) and "parsed" (i.e., analyzed for grammatical structure) (e.g., BNC).

The potential of corpus techniques for investigating patterns of lexis, grammar, semantics, pragmatics, and textual features is well established (e.g., Sinclair, 1991; McEnery & Wilson, 1996; Biber et al., 1998; Kennedy, 1998; Hunston, 2002; Stubbs, 2004). Most work in corpus linguistics to date has relied on word frequency lists, which provide criteria upon which to base a search, and keyword in context (KWIC) concordances, the presentation of every instance of a selected word, phrase, or particle in the corpus down the middle of the page, with a limited amount of cotext on either side. Figure 19.1 provides part of a concordance for the word *meaning*.

The power of the corpus approach lies in the combination of frequency data regarding all the words in a corpus and the verbal environment in which these words occur. This combination permits the detailed investigation of typical patterns of use of lexis and grammar – information which can be obtained at the click of a mouse. A concordance output may appear to be a reified object, but this is not the case, because it may be ordered in various ways from left and right of the keyword or phrase and these changes reveal different collocational and grammatical patterns. Some critics have complained that concordance lines provide no information about situational context. This is also indeed accepted by proponents (e.g., Sinclair, 1991, p. 34; McEnery & Wilson, 1996, p. 79). However, it may be pointed out that situational information can be built into or accompany a corpus and that there is no reason why corpus evidence may not be supported by ethnographic investigation (L. Flowerdew, 2005, 2008). On the other hand, as Stubbs (2004, p. 108) notes, practice has demonstrated that the meaning of a

dictionary like we all do and the first meaning in the dictionary is all to do with #
a proper Italian pronuncation for that meaning leaf or sheet and so really in its basi
heet and so really in its basic basic # meaning portfolio means a case for carrying loo
ussed before she'd also assessed orally meaning they want to check if her teeth are oka
ather different it has a very different meaning and that is that you can't override
tes following a step input oscillates meaning it it goes backwards and forwards and that
o push out the residual demand curve meaning it would comm not only command a
e static properties of the equilibrium meaning we shifted one of the curves and look
kets i'm going to use the word bundles meaning a collection of or a combination of tw
y're falling apart i think what you're meaning is that you've got an i # an interfacia
and it fits the syntax and it fits the meaning that's it i'll # i'll go for that hy
defined concepts which we all know the meaning of but don't need or not even able to d
pecies and i take you understand the meaning of those numbers here minus-two m
index and a little label at the top N meaning the time index so that's the times th
en out # and that's the same same meaning the reason the word's choosed in this
holism bec # for our intentionality meaning our human concepts our human understan
e that it has managed to bleed all the meaning out of maternity we don't have matria
but you under what you understand the meaning of what i'm saying here no i don't
but in economics it has a very specific meaning input-output models were invented by
er goods yeah but this is the literal meaning of this if you increase final demand
e in one country and have very little meaning in another but it has # been a key i
gn so we must talk about the government meaning the legislature and the executive the
u're part of what this can be taken as meaning is that you can consent to be under the
solution but it's a compromise solution meaning that each party will have to give up q
ed of snakes that's F-L-plus S-N-minus meaning this model monkey was afraid of flow
nitive case possessive case of a word meaning kind a kind of something right a nou
about three-thousand years ago # word meaning exactly the same thing is like this

Figure 19.1 Concordance of the word *meaning* from a corpus of academic lectures

search word or phrase is often identifiable within a short span of cotext, enough to fit into one line on the computer screen. Furthermore, most concordancers allow the user to inspect the wider cotext of a selected instance, so the analyst is not limited to the single corpus line.

As already mentioned, corpus techniques have created new knowledge about the behaviour of lexis, grammar, semantics, pragmatics, and textual features. Because corpus linguistics is based on the theory that language varies according to context – across space and time – the potential for finding out new facts about language is infinite. If this theoretical insight is applied to pedagogy, then the case for the use of corpora in teaching becomes very powerful. Because no dictionary or grammar is able to fully describe the language, the educationist, whether materials designer or classroom practitioner – or indeed learners themselves – may play an important role in identifying regularities in the language which are not to be found in such texts.

Another proven benefit of corpus-based approaches to the study of language is that analysis is based on empirical, as opposed to introspected or elicited, data, "real" language as many proponents refer to it.[2] As Aston (2001, pp. 7–8) has pointed out (see also J. Flowerdew, 1996), native-speaker intuition about language is often wrong – on the one hand, many uses included in traditional descriptions

do not occur with any frequency in large general corpora and, on the other hand, many uses which occur in corpus data are not recognized in traditional descriptions. This means that teachers and learners have been being given inaccurate and incomplete descriptions of the language.

What Information Can the Corpus Provide?

A corpus can yield various types of information which can be of potential use in language pedagogy. It can provide information about the behavior of words, multi-word phrases, grammatical patterns, semantic and pragmatic features, and textual properties. Knowledge of these features and their relative frequencies can be helpful to language practitioners in deciding what items to teach and when to teach them, as well as, importantly, providing input for reference materials. In this section various concepts regarding different aspects of language behavior will be presented, each followed by an indication of how the concept might be applied in language pedagogy.

Word frequency

Concept

At the most basic level, the corpus can provide word lists organized either according to frequency or alphabetically. Used in conjunction with the concordancer, frequency is not limited to the word forms, but may extend to the different meanings of a given word or phrase; the editing function of the concordancer can be used to group the items according to the different meanings. Frequency data can also be obtained for recurrent sequences (variously referred to as n-grams, pre-fabs, and lexical bundles) e.g., *I don't know*, *all of a sudden*, *all over the place*, *don't have a clue*. Furthermore, relative frequencies between two or more corpora can be calculated, those words occurring significantly more frequently in one corpus than another being referred to as keywords (Scott & Tribble, 2006).

Application

Frequency information is immensely useful in helping to prioritze what to teach. Aston (2001, p. 8) quite rightly mentions other relevant criteria: range, availability, coverage, learnability, and prototypicality (see also Widdowson, 2004, p. 87; Kaltenböck & Mehlmauer-Larcher, 2005, p. 77), but, other things being equal, frequency of occurrence is an important criterion for syllabus design and teaching. A considerable time ago Nattinger and de Carrico (1992) recommended the application of lexical phrases to language teaching. If one takes the view that all language teaching is LSP teaching, insofar as learners need to acquire a range of registers and genres, then comparative data, as provided by keyword analysis, will provide information regarding what to teach and when to teach in relation to specific genres and registers (see Scott & Tribble, 2006).

Collocation

Concept

Collocation is concerned with how words typically occur (or do not occur) together. Recurrent patterns highlighted by the concordancer will indicate typical collocations, although programs can provide lists of collocates. Hunston (2002, p. 12) gives the example of *shed*, which collocates with *light*, *tears*, *garden*, *jobs*, *blood*, *cents*, *image*, *pounds*, *staff*, *skin*, and *clothes*. Typically different collocates will affect the precise meaning of the word, e.g., *shed blood* means to suffer, *shed pounds* means to lose weight, and *shed image* means a deliberate changing of how one is perceived (Hunston, 2002, p. 12).

Application

How does one correct the learner who says "I will open the air-conditioner," where the collocates are not appropriate? One way, of course is to explain that in standard English one says "switch on" rather than "open" when referring to an air-conditioner or other electrical appliance. However, this lesson is likely to be more powerful and therefore more effective if, instead, the learner can look at concordances of the word *open* + noun phrase and see that while *open* collocates with other concrete nouns such as *gate*, *door*, and *window*, there are no instances of *open* + *air-conditioner*. On the other hand, a concordance of *air-conditioner* will probably yield numerous examples with *switch on* and *switch off*. In addition to this use with students, the concordancer can also give confidence to teachers, especially to less proficient non-native speakers, where they are unsure of their intuitions (J. Flowerdew, 1996).

Colligation

Concept

A distinction can be made between collocation, which is the combination of individual words, and colligation, which refers to how lexical words are associated with particular grammatical words or categories. Hunston (2002, p. 13), again, gives the example of the word *head* which has the following colligations: *of*, *over*, *on*, *back*, and *off*. Again the colligations affect the meaning of the word, thus Hunston gives examples such as *head of department*, *hit someone over the head*, *throw one's head back*.

Application

Kaltenböck and Mehlmauer-Larcher (2005, pp. 73–4) give some examples of pedagogical activities designed to develop colligational awareness. For example, in one simple task students are given a set of sentences where deleted prepositions have to be inserted after searching a corpus:

The building is adjacent . . . the train station.
It is usually a good idea to abide . . . the law.
You should give clear indication . . . your intentions.
He was aghast . . . the violence he witnessed.

In another example task (with the International Corpus of English (ICE) GB corpus, which is tagged for parts of speech), students consider verb complementation with the gerund and the infinitive:

TASK: What can the corpus tell us about the difference in meaning/use between *remember doing something* and *remember to do something?*) (Try [searching for]: "remember to" and "remember" <V> = *remember* followed by a verb.

Kaltenböck and Mehlmauer-Larcher (2005) make the important point that corpus queries such as those required for these tasks require less "expert knowledge" than would be needed if a reference grammar were used, with the knowledge of grammatical metalanguage that would be implied for the latter task.

Semantic preference

Concept

Here we are concerned with how a word or phrase relates to a group of collocating words that (1) share an element of meaning, (2) are related to particular genres or registers, or (3) belong to lexical sets in terms of synonymy, meronymy, antonymy, etc. Semantic preference is arrived at by sorting collocates into groups based on semantic relations such as those just mentioned. The specific semantic preference is labelled by a gloss, such as "words or phrases relating to measurement," "words or phrases belonging to the register of production engineering," or "words or phrases relating to history."

Application

Semantic field theory, which can be seen as an introspective precursor of semantic preference, has been applied (mostly intuitively) in language teaching for a very long time (Corder, 1973, p. 316). Indeed it can be seen as closely related to situational ("at the post office," "at the airport," "in the supermarket," "in the office," etc.) and topical ("travel," "shopping," "family," etc.) syllabuses. It is also implicitly applied in notional syllabuses (Wilkins, 1972). The assumption here is that certain lexical (and grammatical) items belonging typically in given fields are likely to co-occur and can be learned together in semantic sets. However, a corpus approach takes us beyond introspection to identify empirically established relationships. The choice of corpus here is crucial, larger corpora being more reliable, because smaller corpora will not be likely to provide enough data to determine general preferences. On the other hand, specialist corpora consisting of specific genres or registers have great potential for application to LSP.

Semantic prosody

Concept

If semantic preference can tell us about the semantics of a word or phrase, "semantic prosody" (Sinclair, 1991; Louw, 1993), or for Stubbs (2001) "discourse prosody," can tell us about typical pragmatic values – the attitude or evaluation a speaker or writer attaches to what they are saying. Semantic prosody is similar to connotation. However, it does not just apply to a single word, but to the word and its association with its collocates. Thus, to take an example from Stubbs (1996), the word *cause* typically collocates with negatively loaded words – e.g., *accident*, *concern*, *damage*, *death*, *trouble* – and thereby takes on a negative semantic prosody; p*rovide*, on the other hand, is typically used with positive collocates – e.g., *aid*, *care*, *food*, *opportunities*, *relief*, *support* – and thus takes on a positive semantic prosody. Most studies of semantic prosody describe examples in simple terms of positive or negative evaluation. However, it seems that finer grained analysis is possible. Thus Hunston (2002, p. 141) gives the example of *sit through*, which is often used with lexical items which indicate boring or lengthy things.

Application

The analysis of the semantic prosodies associated with the lexical items in a corpus is a way to acquire context knowledge which is important for writers trying to master tasks within a specific genre (Tribble, 2000). This sort of information is now starting to be incorporated into dictionaries, but a learning activity in the form of analyzing words in context and identifying their semantic prosodies might be a more effective learning strategy, insofar as learners are more likely to remember what they themselves have discovered.[3]

Register and genre

Concept

Research in corpus linguistics has done much to show how patterns may vary across various registers or genres. As Biber and Conrad (2001, p. 332) put it, "strong patterns of use in one register often represent only weak patterns in other registers." To illustrate this, Biber and Conrad show, for example, how the 12 most frequent lexical verbs (*say*, *get*, *go*, *know*, *think*, *see*, *make*, *come*, *take*, *want*, *give*, and *mean*) in a corpus of 20 million words drawn from four registers (conversation, fiction, newspaper language, and academic prose) are very unequally distributed across the four registers. These verbs, for example, represent 45 percent of all verbs in conversation versus only 11 percent for academic prose.

Application

Biber and Conrad (2001) argue that the verbs referred to above should be given priority in pedagogy. In practice, however, they note that low-level ESL grammar books tend not to use these verbs, preferring activity verbs such as *eat*, *play*, *work*,

run, *travel*, and *study*, which, as they concede in a footnote, are easier to learn. Nevertheless, Biber and Conrad argue that just because they are more difficult to learn does not mean that the high-frequency verbs should be neglected, "as these are the ones students will most often hear in their day-to-day interactions with native speakers" (p. 333).[4]

At a more micro level, working with small corpora composed of specific text types, Gavioli (2001) has shown how particular recurrent patterns tend to occur within such corpora. Comparing two corpora, one composed of lonely hearts ads and the other of letters to a newspaper agony aunt, for example, she (or, in fact, her students, because in her paper Gavioli is showing how corpus analysis can be done by learners (see below)) shows certain similarities and differences. Taking the adjectives *pretty*, *attractive*, and *beautiful*, for example, she shows that *pretty* and *attractive* always refer to people's physical appearance in both corpora. *Beautiful*, in the letters, however, also refers to music and the home. In addition, neither *pretty* nor *attractive* occurs in a series of adjectives. However, *beautiful* occurs in a co-ordinate pattern with another positive adjective, in phrases such as *mature and beautiful*; *beautiful and well-behaved*; *beautiful and wonderful*; *sweet and beautiful*. This is not the sort of information that can be found in reference grammars or dictionaries and it provides a strong argument for a corpus-based approach to the development of genre awareness (J. Flowerdew, 1993b).

What Corpora?

One of the problems with applying corpora to language teaching is deciding which the most appropriate corpora are. As Leech (1997, p. 18) has pointed out, "the corpora which are easiest to compile are not necessarily those which are most useful for language learning purposes." Not all corpora will be suitable for all learners.

Until recently, the most pressing problem in this area was the dearth of spoken corpora, most corpora being wholly or primarily made up of written language. The reason for this is simple. It is difficult and expensive to collect spoken language, which then has to be recorded and transcribed. It is true that spoken corpora are starting to be created – for example the CANCODE corpus of spoken English developed jointly by the University of Nottingham and Cambridge University Press – but there is still an emphasis on the written word (not to mention problems of accessibility). The BNC, for example, has 90 million written words compared to 10 million of speech). Given the emphasis in modern-day language pedagogy on the spoken word, this is a serious problem.

Another problem is that most corpora are based on native-speaker models. In a climate where there is much discussion of the role of world Englishes in language pedagogy, the use of native-speaker models may be questioned (Hunston, 2002). This does not just concern lexico-grammar. As Carter (1998b, p. 49) has demonstrated, colloquial speech is deeply embedded in cultural understandings. The simplest of phrases may require knowledge of the culture for understanding.

Among many examples, Carter provides the following service encounter from CANCODE:

[In a fish and chip shop]
A: Can I have chips, beans, and a sausage?
B: Chips, beans, and a sausage.
A: Yeah.
B: Wrapped up?
A: Open please.

Carter points out how in this extract the word "open" in this particular context is used as an antonym of "wrapped up" and "carries a specific cultural meaning of food being served in paper so that it can be eaten immediately, even perhaps while walking home" (p. 48). Carter asks to what extent such cultural allusions should be removed. Furthermore, he asks how relevant it is to be able to make observations such as that fish and chip shops in Britain serve not only fish and chips, but also other food, such as sausages, burgers, and curry.

The foregoing suggests that corpora made up of different language varieties might be needed. Hong Kong learners or Filipino learners, it might be argued, should have as their target educated Hong Kong or educated Filipino English, not British English. Similarly, it would seem sensible that learners of French in Canada might want a standard and hence a corpus based on Canadian French rather than the metropolitan variety. The problem is being addressed to a degree, for English, with the ICE corpora, referred to earlier, a suite of corpora of 15 different national/regional varieties, such as Australian English, British English, East African English, Filipino English, Indian English, etc. Given the complexity of coordinating and collecting such a range of corpora, however, it is perhaps understandable that these corpora are relatively small, at one million words each. Of course, the question of what standard to adopt is itself controversial. To take the example of Canadian French, many learners want to acquire the metropolitan standard, even though they will be using their French in Canada. To take another example, at a recent corpus conference an Indian member of the audience was asked if Indians would want to learn Indian English. His answer was that they would definitely not want to be associated with such a variety which they did not even acknowledge as such, preferring so-called "standard" English. This raises the question of language rights, in this case the rights of the learners (or their parents) to have the target variety that they want. In addition, where regional or local varieties have developed, in a globalizing world, with all that goes with it – mass migration, mass tourism, international business travel, the internationalization of (especially tertiary) education, use of the Internet and other electronic communication devices, and the internationalization of popular culture and mass media – learners may need not only the local variety, but also some standard for international intelligibility.

An alternative solution in terms of appropriate models might lie in lingua franca corpora, corpora composed of language produced by proficient non-

native speakers who are interacting with each other or with native speakers. In English, it is said that more English is spoken in the world among non-natives than natives, so lingua franca corpora might seem a logical way to go. However, research to date has not come up with systematic descriptions of such language and it is questionable whether – certainly at the level of phraseology and the grammatical code – such systematic patterns are discernible. Interestingly Anna Mauranen (2006; personal communication June 2006) has identified in her ELFA (English as a Lingua Franca for Academic Purposes) corpus certain pragmatic regularities, such as greater use of grammatical rephrasing and a greater tolerance for ambiguity. But she has not identified any new lexico-grammatical or phraseological regularities.

Further confusing the picture as regards suitable corpora, there are other learner differences that need to be taken into account. For example, a model for young learners might be child language, teenagers may want a teenage model, women and men might want models of the speech of either gender; then again, learners may want specific academic or professional language (see below on corpora and LSP). It is true that there are different types of corpora or sub-corpora (for example, the CHILDES corpus of children's speech (MacWhinney & Snow, 1991) and the British National Corpus has a section on young people's spoken language, referred to separately as the COLT corpus (Stenström et al., 2002)). However, these corpora or sub-corpora have not been designed with language teaching specifically in mind and their suitability, certainly in terms of their size and representativeness, might be questioned.[5]

Finally, the authenticity of corpus data may mean that it is difficult for less advanced learners to process. Perhaps corpora of simplified language might be needed for such learners, or some sort of filter which removes concordances which contain vocabulary items which do not occur in a pre-established list (Kuo et al., 2001).

Applications

Applications of corpus linguistics to language teaching may be direct or indirect (Stubbs, 2004). A direct application would be advanced users of academic English using a corpus of the language of their speciality to assist them in writing academic papers (see Lee and Swales, 2006 for an account of such a procedure). Indirect applications would be the application of corpus findings to the creation or refinement of dictionaries, reference grammars, and pedagogic materials.

Indirect applications

Use in developing reference material

One of the first applications in this area was *Collins COBUILD English Language Dictionary* (1987) edited by John Sinclair; Other dictionaries have made use of

corpora to a greater or lesser extent, e.g., *Longman Dictionary of Contemporary English*, *Oxford Advanced Learner's Dictionary*, *Macmillan English Dictionary for Advanced Learners*. As Leech (1997, p. 14) points out, some of the advantages of corpus-based lexicography are that corpus data:

- can be searched quickly and exhaustively,
- can provide frequency data,
- can be easily processed to produce updated lists of words,
- can provide authentic examples for citation,
- can readily be used by lexicographical teams for updating and verifying other levels of descriptions such as dictionary definitions.

A precursor of grammars totally based on corpus data was *A Comphrehensive Grammar of the English* Language (Quirk et al., 1985), which relied on manually collected examples of use that were stored in a giant database. This might be called a *corpus-informed* grammar. The first grammar to be fully based on a corpus, what might be called a *corpus-driven grammar*, was *Collins COBUILD English Grammar* (Sinclair, 1990). This has been followed by the *Longman Grammar of Spoken and Written English* (Biber et al., 1999), and, more recently, by the *Cambridge Grammar of English* (Carter & McCarthy, 2006). At this point, it might be noted that, while the *Longman Grammar* is no doubt a great achievement, especially in the great advance achieved in the incorporation of frequency data according to four different domains of use, the corpus-driven grammars, especially COBUILD and Cambridge tend to be inconsistent in their coverage of the basic features of the language. In terms of comprehensiveness, Quirk et al. (1985) cannot be beaten. No doubt with further work the comprehensiveness of corpus-driven grammars will improve.

Pedagogic materials

Again, Collins (now HarperCollins), with John Sinclair as editor in chief, were the first here, with an extensive series called *Collins COBUILD English Guides*. Titles focused on such linguistic features as *Determiners* (Berry, 1996) *Linking Words* (Chalker, 1996), and *Reporting* (Thompson, 1993). Coming again out of the work at Birmingham was the first proposal for syllabus design to be based on corpora – *The Lexical Syllabus: A New Approach to Language Teaching* (D. Willis, 1990), and also *Collins COBUILD English Course* (Willis & Willis, 1989).

Direct Applications

Corpora and syllabus design

If one accepts a corpus view of language, i.e., that it consists to a great extent of recurrent patterns (what Sinclair, 1991 refers to as the "idiom principle"), then important implications apply for syllabus design. Instead of being organized in

terms of grammatical forms, the syllabus can be designed around the most important recurrent patterns (see Sinclair & Renouf, 1988; Willis, 1990; Willis & Willis, 1989). This type of syllabus is referred to as a lexical syllabus, although this is somewhat misleading, as it is designed around lexical patterns, not single words. The idea of basing a syllabus on patterns of use was, in fact, put forward as early as 1980 by Nattinger:

> Perhaps we should base our teaching on the assumption that, for a great deal of the time anyway, language production consists of piecing together the ready-made units appropriate for a particular situation and that comprehension relies on knowing which of these patterns to predict in these situations. Our teaching, therefore, would center on these patterns and the ways they can be pieced together, along with the ways they vary and the situations in which they occur. (cited in Richards & Rogers, 2001, pp. 133–4)

Although the emphasis is on lexical patterning in the lexical syllabus, grammar is not neglected, it can be argued, as the main lexical patterns will incorporate the main grammatical forms. Willis (1990, p. vi) takes this a stage further, claiming that "the lexical syllabus not only subsumes a structural syllabus, it also indicates how the structures which make up the syllabus should be exemplified." For the COBUILD course, for the first level, the most frequent 700 words were selected from the COBUILD corpus, these words accounting, according to Willis (1990, p. vi), for around 70 percent of all English text.

The underlying principle of the lexical syllabus is frequency. Sinclair and Renouf (1988) argue that the most frequent words typically have a range of uses and that many of these uses are typically not covered in beginners' courses. They give the example of the word *make*. This word most typically occurs in patterns such as *make decisions*, *make discoveries*, *make arrangements*. These abstract uses are more frequent than the concrete use, as in *make a cake*. Sinclair and Renouf thus argue that the abstract forms should be taught to beginners, as well as the concrete ones. It should be noted, however, that this is a very strong argument, which neglects the counter-argument of "teachability," the fact that the concrete meaning is more easily taught. Widdowson (2004, p. 87) puts forward the teachability argument, as follows: "Words and structures might be identified as 'pedagogically' core or nuclear, and preferred as a prototype at a particular learning stage because of their coverage or their generative value, because they are catalysts which activated the learning process, whatever their status might be in respect to their actual occurrence in contexts of use." This is the same argument as noted with the examples given by Biber and Conrad of the most frequent verbs. On the other hand, it might be argued that the teachability of the concrete forms should imply that they are taught first, but then the other abstract uses immediately follow. This seems to have been the principle adopted in the COBUILD dictionary, where concrete meanings still come before abstract ones. Thus the concrete meaning of "lifebelt," for example, comes before the metaphorical ones.

Data-driven learning

The beginning of classroom concordancing is generally attributed to Tim Johns and his work with non-native speaking postgraduate science students at the University of Birmingham. Johns referred to this approach as "data-driven learning" (DDL). In this form of learning, learners are seen as "language detectives" (Johns, 1997, p. 101), seeking answers to questions that can be found by means of corpus queries and/or concordance lines. Learners are detectives because concordances do not offer explanations; they simply provide (patterned) data for analysis. Learners are required to identify and analyze the recurrent patterns to be found in the concordance lines and make their own generalizations. They may do this by working on the concordance print-outs (Johns' preferred method) or directly with the computer and the corpus. Johns (cited in Ma, 1994, p. 197) sees this approach as falling between "the highly-organized, graded and idealized language of the typical coursebook" and the "potentially confusing but far richer and more revealing 'full flood' of authentic communication."

The DDL approach has alternatively been described as one of "authenticity and discovery" (Ma, 1994, p. 197). Leech (1997, p. 3) comments how "it often happens that a student working on a relatively small corpus assignment comes up with original observations and discoveries which have probably never been brought to notice before, even in the most detailed dictionaries and grammars of language." Indeed, an example of such a discovery has already been presented in this chapter with Gavioli's example of the genre-specific use of particular adjectives in letters to agony aunts and lonely hearts ads. A number of writers provide examples of the sort of activities that can be exploited in DDL (e.g., Aston, 2001; Dodd, 1997; Gavioli, 2001; Hunston, 2002) A commercially produced workbook based on concordance print-outs is Thurston and Candlin (1993). Here is just one simple activity from Gavioli (2001, p. 125):

WORKSHEET

1. In groups of two or three use the corpus of Lonely hearts ads to identify 4 or 5 patterns which are typically used to do each of the following
   - introduce the seeker
   - add descriptions of the seeker
   - introduce the sought
   - add descriptions of the sought

   You can generate concordances to get suggestions.

2. Use the patterns you have identified to produce at least two new lonely hearts ads. Complete any missing parts of the text as necessary.

This activity is interesting because it requires the learners to go beyond just analyzing concordance lines, i.e., working out possible functions for patterns which are presented to them. Here they first have to come up with their own corpus queries and then find patterns which typically realize the functions which

are embodied in their queries. This is, therefore, the reverse procedure to the more typical one.

This activity takes us on to ones used with more advanced learners, where they use specialist corpora for real-world writing tasks. In probably the first example of such an application, Ma (1994) created a corpus of computer user manuals for his students, who had to write a chapter of such a manual, using the corpus and a concordancing program as a writing resource. In a similar approach, Bianchi and Pazzaglia (2007) created a corpus for psychology students consisting of experimental articles in that discipline, the task being for students to write a research article of their own, using the corpus as a resource. Interestingly, this corpus was divided into sub-corpora of the different "moves" (Swales, 1990) in the articles. Taking this sort of procedure a stage even further, Lee and Swales (2006) had a heterogeneous group of graduate students who created their own corpora specific to their particular discipline. These corpora were used as a resource for working on the writing required on their higher degree programs. One issue that has to be confronted with this type of application is where to draw the borderline between reusing typical phraseological patterns, on the one hand, and copying longer sections of texts extracted from the corpus – a case of plagiarism – on the other (see Pecorari, 2003, 2006; J. Flowerdew & Li, 2009). This issue needs to be handled carefully by the teacher, although it could also be seen as an opportunity to alert students to issues of plagiarism.

Corpora and LSP

The three studies reported in the preceding paragraph are all concerned with LSP and they are very recent. This section will step back from these studies and consider corpora and LSP since its beginning, because LSP (primarily English for specific purposes (ESP)) has been one of the major fields of development for corpus application to language teaching (see Gavioli, 2005 for a complete monograph on this topic, and L. Flowerdew, 2002 for an overview paper for English for academic purposes (EAP) – here we cannot do full justice to this area, merely citing a number of key examples). The earliest known concordances, which, of course, were compiled by hand, were based on biblical texts. The first complete concordance of the Latin Bible was created by the Benedictine Hugo de San Charo in the thirteenth century (Tribble & Jones, 1990, p. 7). Early concordances, in their focus on one register (biblical texts) can, in fact be seen as a form of LSP. In the modern era, a precursor of modern corpus work is an influential article by Barber (1962). Barber's study, carried out by hand, and with a view of pedagogic application, is based on three scientific research articles, making a "corpus" of about 23,000 words (tiny by today's standards). Barber studied various syntactic features of this corpus, one of the most striking findings being the very low occurrence of progressive aspect. This finding, as Swales (1988, p. 1) reports, was influential in the teaching of English for Science and Technology (EST), suggesting that attention should be given to the other aspects rather than the progressive.

Probably the earliest application of corpus analysis to syllabus and materials design in ESP was conducted by J. Flowerdew (1993a), where a corpus was created from transcriptions of lectures that biology students concurrently taking an ESP course attend during their biology course, along with the assigned readings. This corpus was used as the basis for selecting key vocabulary (about 1,000 items, chosen on the basis that they occur more than 10 times in the corpus). This lexis formed the core vocabulary of the ESP course. With the use of concordances, the most typical recurrent phraseological patterns in which this vocabulary occur were identified and incorporated into materials. In addition, based on a close reading of the text and examination of the word list and concordancing lines, important notions and functions were identified, along with their typical realization patterns. These were also incorporated into the syllabus and course materials. This same database was used when it came to assessing the performance of students during and after the course.

Contemporary work in ESP is very much influenced by genre theory (Swales, 1990; Bhatia, 1993) and ESP corpora are typically compiled from texts belonging to the same genre. The emphasis on genre has meant a move away from a simple focus on the frequency of syntactic structures to a consideration of form–function relations and how linguistic features correlate with generic moves and other pragmatic aspects of particular genres. A good example of the genre approach is Gledhill's (2000) study of research articles in the field of cancer (also summarized in Gavioli, 2005, p. 57). The corpus for this work is divided into sub-corpora according to the generic moves of research articles (introduction, method, results, discussion, conclusion). This allows Gledhill to identify salient lexis in the various moves. This is done using key-word analysis, key-words being those words which are statistically more frequent in one move than in the rest of the corpus.[6] Having identified the key-words, Gledhill then uses concordances to identify their particular rhetorical functions e.g., "such" is often used to reformulate biochemical processes (e.g., *antitumour agents such as NMY*; *use of hormonal enzymes such as dismutase*), "can" is used to express potential clinical processes (e.g., *methods can be considered*; *alterations can be prepared*). Finally Gledhill shows how the key-words often form part of recurrent phraseological patterns. So, for example, "to" is found in constructions which take the following pattern [biomedical process] (possessive) <ability to> [biochemical process], as in expressions such as:

[the reactant] Its	*ability to*	*alter tolerance to self*
we extended its [tumor]	*ability to*	*differentiate*
calibrating their [leukocytes]	*ability to*	*modify factor specific DNA*

In another example of genre-based corpus work, Swales et al. (1998) focused on the use of the imperative in research articles. Using a corpus of 50 articles (5 from each of 10 disciplines), computer-assisted analysis provided data on all instances of imperatives in both the main texts and footnotes of these articles. In those disciplines where imperatives occurred in the main text (only five out of the ten selected), they tend to occur in the more argumentative sections of

articles, but are very unevenly distributed across disciplines, being most prevalent in fields where mathematical reasoning occurs. In addition, there are a range of field-specific usages. In terms of function, based on interviews with some of the writers, Swales et al. argue that the decision to use imperatives may be part of a bundle of grammatical features that promote irony, closer collegiality, or playfulness, which they hypothesize is a contemporary trend of scholarly writing in the post-modern age.[7] Imperatives are not used in the research article as face-threatening devices, as suggested in standard grammars, but as one of the resources available to writers which allows them to maintain a harmonious relationship with their readers.

The use of computers is also helpful for taxonomic research into the various functional categories which a given linguistic form is used to realize. In order to develop a taxonomy of the functions of reporting verbs in academic articles, for example, Hyland (2002) used the following procedure. First, he computer-searched his corpus for canonical citation forms, such as a date in brackets, a number in squared brackets, and Latinate references to other citations. In addition, a concordance was made of all of the names in the bibliographies of the articles which made up the corpus, and of second person pronouns. This search yielded all of the citations in the corpus, from which Hyland was able to extract the reporting verbs and classify them according to the specific type of activity they refer to: research acts (which contain verbs which represent experimental activities or actions carried out in the real world (e.g., *observe*, *discover*, *notice*, *show*)), cognition acts (which contain verbs which are concerned with the researcher's mental processes (e.g., *believe*, *conceptualize*, *suspect*, *assume*, *view*)), and discourse acts (which contain verbs which involve linguistic activities and focus on the verbal expression of cognitive or research activities (e.g., *ascribe*, *discuss*, *hypothesize*, *report*, *state*)).

The sort of analyses described here might seem very fine-grained, but in ESP situations they have the potential to provide teachers with very specific information which can be incorporated into teaching. At the level of the classroom, an interesting teaching procedure is reported by Weber (2001). Weber took "a concordance- and genre-based" (Weber, 2001, p. 14) approach to teaching law students to write academic essays. Students first of all had to analyze a corpus of essays written by native-speaker law students, either individually or in small groups, and identify, through consensus, elements which seemed essential to the structure of the essays (they identified four: identifying and/or delimiting the legal principle; referring to the authorities; applying these judicial precedents and/or reasoning on the basis of these precedents; moving toward a conclusion and/or giving advice to the parties concerned). The students then had to search for language that seemed to correlate with the structural elements they had identified. In groups, they selected the most significant examples. They then were given the opportunity to work with different corpora of non-legal genres, searching for the language they had identified in the legal genre and seeing how different the uses were in the other genres. Next, they were given case studies and asked to write very short essays, incorporating the four structural elements they had identified and using the language they had identified in the concordances. The essays were then subjected to peer review and group discussion and finally

a short conference was held with each student. Weber (2001, p. 19) describes the activity as giving the students "a firm foundation both in essay writing and in legal reasoning." It is also a very good example of how data-driven learning can be applied using a communicative, task-based format.

While the examples cited thus far in the field of ESP have been quite specific, mention should be made of more generic EAP (English for Academic Purposes) work. One example here would be Coxhead's Academic Word List (Coxhead, 2000; Coxhead & Nation, 2001). This is a list of 2,000 words which are identified on the basis of their frequency and range in academic English. The list is derived from a comparison between a 3.5 million-word corpus of academic English (from different disciplines and genres) and a reference corpus consisting of the same amount of fictional writing. The words included in the list are neither the typically very high frequency words of everyday English nor the technical language of specific disciplines. Also worthy of mention are large-scale academic corpora which are being made available in the public domain. The best-known of these is the MICASE corpus of American spoken academic language developed at the University of Michigan (www.lsa.umich.edu/eli/micase/index.htm), but there is also the counterpart BASE corpus of spoken British academic language developed by the universities of Reading and Warwick (www.rdg.ac.uk/AcaDepts/ll/base_corpus/) and its companion written BAWE corpus (www.coventry.ac.uk/researchnet/d/505/a/2850).

Learner Corpora

A learner corpus is a collection of texts which have been produced by learners of a language. Learner corpora allow for the comparison of learner language with native-speakers of the target language (L2 vs. L1) or with other groups of learners (L2 vs. L2) (see Granger, 2004a for an overview). The best-known work in this area has been conducted by Sylviane Granger and her colleagues at the University of Louvain in Belgium. This work has resulted in the creation of a suite of different, but comparable, learner corpora of argumentative essays from a whole range of different L1 learners. Corpora such as those contained in ICLE can be used as a tool for contrastive interlanguage analysis and for error analysis (Granger, 2004b). As well as identifying discrepancies between different stages of interlanguage and between L2 and L1 users on the basis of "errors," frequency data drawn from learner corpora can show how learners may over- or under-use certain patterns of the target language, features which have been identified as the reason for LS speakers from sounding "non-nativelike."[8] In terms of the quantity of work being done with learner corpora, this is very popular. At the time of writing. the online bibliography for learner corpora at the Centre for English Corpus Linguistics, Université catholique de Louvain (Belgium) (http://cecl.fltr.ucl.ac.be/learner%20corpus%20bibliography.html) had over 330 entries.

In terms of application, the contribution of learner corpora to language teaching is primarily "indirect," to use Stubbs' (2004) term; learner corpora have primarily been used to assist in the production of reference tools. Granger (2004a, 2004b)

cites a number of dictionaries which have made use of learner corpus findings. Two examples of these are the *Longman Dictionary of Contemporary English* (using a 10 million-word learner corpus), and the *Cambridge Advanced Learner's Dictionary* (using a 16 million-word corpus), both of which contain notes drawn from analysis of their respective corpora advising users on how to avoid common errors. Granger (2004b, p. 3) notes that information is generic in nature (i.e., not specific to any particular L1), due to the limited space available in paper dictionaries. With the greater use of online dictionaries it is likely that information for specific L1 groups will be made available.

Concerning more "direct" uses, Granger (2004a, 2004b) cites a number of computer programs that are designed to help learners with the types of errors which have been identified in learner corpora. One well-known example of these is Milton's (1998) *Word Pilot*, a program which allows students and teachers to explore lexical patterns in any text type, using lists of problematic words and phrases identified by analysis of a large corpus of Cantonese mother tongue users of English. Target lexis extracted by analysis of other learner corpora can be loaded into the program as well. Milton (2006) has since extended this development to a suite of programs that assist second language writers to improve their written fluency and proofread for common errors. These programs include online vocabulary databases, an Internet-based grammar, various lookup resources, and a tool that assists teachers to insert feedback in students' electronic documents (Milton, 2006). This suite of programs can be accessed at http://mywords.ust.hk.

Another example of the application of learner corpora is *IWiLL* (Wible et al., 2001), an interactive web-based tool which allows students and teachers to create and use an online database of Taiwanese learners' essays and teachers' error annotations. Granger (2004b) also cites an application by Hewings (2000), who used a learner corpus and an ESP corpus to compare a range of linguistic features with his students. One example Hewings gives is of the personal pronoun *I*. It was found that where students tend to use *I* to express their own opinion (*I believe*, *I think*, *I suppose*), in the ESP corpus *I* was mainly used as a text-organizing device or for reporting a procedure (*As I have already pointed out*, *The survey I conducted among my students*).

Conclusion

In this review, an attempt has been made to cover as much ground as possible. Inevitably, however, in a chapter of this type, there remain some gaps. This is due partly to limitations of space, partly the ignorance of the author, and partly because certain issues have not been dealt with in the literature very thoroughly, if at all. This review has taken an overall positive view of corpora and language pedagogy. Perhaps some caveats are in order in this conclusion. Cook (1998) claims that some corpus linguists "overreach" themselves and that "they talk as though the entire study of language can be replaced by the study of their collections" (p. 57). While this may or may not be true (and if it is, it probably only

applies to a very few applied linguists), it *is* true to say that corpus applications to language teaching have been promoted by enthusiasts. However, as indicated already, many more applications are taking place in tertiary institutions (where teachers are better resourced and more research-oriented) than in schools. Ways need to be found to encourage corpus-based work at this level. While, as has been seen, a lot has been written extolling the virtues of corpora in language pedagogy, less has been written about some of the problems, for example some of the difficulties novice corpus users encounter. More needs to be written up on this (although see, e.g., Bernardini, 2000; Frankenberg-Garcia, 2005; Kennedy & Miceli, 2001). Another issue worthy of consideration is the question of the need to keep corpora up to date. Language is changing very quickly, but data for some of the major corpora currently in use were collected a considerable time ago. One can wonder to what extent, for example, the teenage language sub-component of the BNC (the COLT corpus) can still be said to represent the way English is spoken by British teenagers today, given that the data were collected in the mid-1990s. There is a need for new corpora or an updating of existing ones. Other issues worthy of consideration are the relative paucity of literature on teacher education (but see, e.g., Hunston, 1995; Farr, 2008; O'Keeffe & Farr, 2003; Tsui, 2004) and evaluation of pedagogic applications (but see, e.g., Cobb, 1997, 1999; Farr, 2008; Kennedy & Micheli, 2001; Yoon & Hirvela, 2004). Overall, however, the outlook for corpora and language teaching looks healthy. The prognosis, if not perfect, is very good.

NOTES

I should like to acknowledge feedback from David Lee on an earlier draft of this paper.

1 Although, looked at in another way, corpora may contain language which individuals are unlikely to be familiar with, in so far as they may contain language from genres or registers outside a given individual's experience.
2 The term "real" in this context has been critiqued (see e.g., Carter, 1998a and Cook, 1998).
3 This process is known as data-driven learning (see below).
4 The same caveats regarding frequency, as noted above, apply here.
5 It should be noted that CANCODE, which is not available publicly, was developed "with an eye to [the data's] potential relevance to ELT" (Carter, 1998a, p. 43).
6 This represents an advance on the approach adopted by J. Flowerdew (1993a), who used raw frequency data, software not being available at the time to perform key-word analysis. See Scott and Tribble (2006) for a book-length treatment of the application of key-word analysis to language teaching.
7 Swales et al. are not the only corpus-oriented ESP practitioners to make use of specialist informants (e.g., Hyland, 2000; L. Flowerdew, 2005, 2008).
8 Users of learner corpora are open to the criticism that native speakers will be the target norm (Hunston, 2002, pp. 211–12). This may not necessarily be the case, however. It is up to the users of learner corpora to select the norm as they see fit. It is quite possible

to conceive of a lingua franca corpus as the norm, or a corpus of educated Hong Kong or Filipino English, although none have been used as such to date.

REFERENCES

Ahmed, K. & Davies, A. (1997). The role of corpora in studying and promoting Welsh. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (eds.), *Teaching and language corpora* (pp. 157–72). London: Longman.

Aston, G. (2001). Learning with corpora: An overview. In G. Aston (ed.), *Learning with corpora* (pp. 7–45). Houston: Athelstan.

Barber, C. L. (1962/1988). Some measurable characteristics of modern scientific prose. In J. Swales (ed.), *Episodes in ESP* (pp. 3–14). London: Prentice-Hall.

Belz, J. A. (2004). Learner corpus analysis and the development of foreign language proficiency. *System* 32, 4, 577–91.

Bernardini, S. (2000). Systematising serendipity: Proposals for concordancing large corpora with language learners. In L. Burnard & T. McEnery (eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 225–34). Frankfurt: Peter Lang.

Bernardini, S. (2004). Corpora in the classroom. An overview and some reflections on future developments. In J. M. Sinclair (ed.), *How to use corpora in language teaching* (pp. 15–36). Amsterdam/Philadelphia: John Benjamins.

Berry, R. (1996). *Determiners.* London: Collins.

Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings*. Harlow, UK: Longman.

Bianchi, F. & Pazzaglia, R. (2007). Student writing of research articles in a foreign language: Metacognition and corpora. In R. Facchinetti (ed.), *Corpus Linguistics 25 Years on* (pp. 259–87). Amsterdam/New York: Rodopi.

Biber, D. & Conrad, S. (2001). Quantitative corpus-based research in TESOL: Much more than bean counting. *TESOL Quarterly* 35, 2, 331–5.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating grammar of spoken and written English*. London: Longman.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (eds.) (1999). *Longman grammar of spoken and written English.* London: Longman.

Bolly, C. (2005). *Constructions récurrentes, collocations et séquences figées avec le verbe à haute fréquence "prendre": Pour une méthode "mixte" d'analyse de corpus (FL1 et FL2)*. In C. Cosme, C. Gouverneur, F. Meunier, & M. Paquot (eds.), *Proceedings of the Phraseology 2005 Conference* (pp. 57–60), Louvain-la-Neuve, October 13–15.

Carter, R. (1998a). Orders of reality: CANCODE, communication and culture, and Reply to Guy Cook. *ELT Journal* 52, 1, 43–56; 64.

Carter, R. (1998b). Telling tails: Grammar, the spoken language and materials development. In B. Tomlinson (ed.), *Materials development in language teaching* (pp. 67–86). Cambridge: Cambridge University Press.

Carter, R. & McCarthy, M. (eds.) (2006). *The Cambridge Grammar of English*. Cambridge: Cambridge University Press.

Chalker, S. (1996). *Linking words*. London: Collins.

Cobb, T. (1997). Is there any measurable learning from hands on concordancing? *System* 25, 3, 301–15.

Cobb, T. (1999). Breadth and depth of lexical acquisition with hands-on concordancing. *Computer Assisted Language Learning* 12, 345–60.
Cook, G. (1998). The uses of reality: A reply to Ronald Carter. *ELT Journal* 52, 1, 57–63.
Corder, S. P. (1973). *Introducing applied linguistics*. Harmondsworth, UK: Penguin.
Coxhead, A. (2000). A new academic word list. *Tesol Quarterly* 34, 2, 213–38.
Coxhead, A. & Nation, P. (2001). The specialised vocabulary of English for Academic Purposes. In J. Flowerdew & M. Peacock (eds.), *Research perspectives on English for Academic Purposes* (pp. 252–67). Cambridge: Cambridge University Press.
Dodd, B. (1997). Exploiting a corpus of written German for advanced language learning. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (eds.), *Teaching and language corpora* (pp. 131–45). London: Longman.
Farr, F. (2008). Evaluating the use of corpus-based instruction in a language teacher education context: Perspectives from the users. *Language Awareness* 17, 1, 25–43.
Flowerdew, J. (1993a). Concordancing as a tool in course design. *System* 21, 2, 231–44.
Flowerdew, J. (1993b). A process, or educational, approach to the teaching of professional genres. *ELT Journal* 47, 4, 305–16.
Flowerdew, J. (1996). Concordancing in language learning. In M. Pennington (ed.), *The power of CALL* (pp. 97–113). Houston: Athelstan.
Flowerdew, J. & Li, Y.-Y. (2009). Plagiarism and second language writing in an electronic age. *Annual Review of Applied Linguistics* 27, 161–83.
Flowerdew, L. (2002). Corpus-based Analyses in EAP. In J. Flowerdew (ed.), *Academic discourse* (pp. 95–114). London: Longman.
Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus-based methodologies. *English for Specific Purposes* 24, 321–32.
Flowerdew, L. (2008). Corpora and context in professional writing. In V. Bhatia, J. Flowerdew, & R. Jones (eds.), *Advances in discourse studies* (pp. 115–27). London: Longman.
Frankenberg-Garcia, A. (2005). A Peek into what today's language learners as researchers actually do. *International Journal of Lexicography* 18, 3, 335–55.
Gavioli, L. (2001). The learner as researcher: Introducing corpus concordancing in the classroom. In G. Aston (ed.), *Learning with corpora* (pp. 108–37). Houston: Athelstan.
Gavioli, L. (2005). *Exploring corpora for ESP learning*. Amsterdam/Philadelphia: John Benjamins.
Gavioli, L. & Aston, G. (2001). Enriching reality: Language corpora in language pedagogy. *ELT Journal* 55, 3, 238–46.
Gledhill, C. (2000). The discourse function of collocation in research article introductions. *English for Specific Purposes* 19, 115–35.
Granger, S. (2004a). Computer learner corpus research: Current status and future prospects. In U. Connor & T. A. Upton (eds.), *Applied corpus linguistics: A multidimensional perspective* (pp. 123–45). Amsterdam: Rodopi.
Granger, S. (2004b). Practical applications of learner corpora. In B. Lewandowska-Tomaszczyk (ed.), *Practical applications in language and computers (PALC 2003)* (pp. 291–301). Frankfurt: Peter Lang.
Hewings, M. (2000). *Using computer-based corpora as a teaching resource*. Available from www.realenglish.tm.fr.
Higgins, J. (1988). *Language, learners and computers*. London: Longman.
Higgins, J. & Johns, T. (1984). *Computers in language learning*. London: Collins.

Hunston, S. (1995). Grammar in teacher education: The role of a corpus. *Language Awareness* 4, 1, 15–31.
Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. Harlow, UK: Pearson Education.
Hyland, K. (2002). Activity and evaluation: Reporting practices in academic writing. In J. Flowerdew (ed.) A*cademic discourse* (pp. 115–30). London: Longman.
Johns, T. (1988). Whence and whither classroom concordancing. In T. Bongaerts, P. De Haan, S. Lobbe, & H. Wekker (eds.), *Computer applications in language learning* (pp. 9–27). Dordrecht: Foris.
Kaltenböck, G. & Mehlmauer-Larcher, B. (2005). Computer corpora and the language classroom: On the potential and limitations of computer corpora in language teaching. *ReCALL* 17, 1, 65–84.
Kennedy, C. & Miceli, T. (2001). An evaluation of intermediate students' approaches to corpus investigation. *Language Learning & Technology* 5, 3, 77–90.
Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Addison Wesley Longman.
King, P. (1997). Creating and processing corpora in Greek and Cyrillic alphabets on the personal computer. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (eds.), *Teaching and language corpora* (pp. 277–91). London: Longman.
Kuo, C.-H., Wible, D., Wang, C.-C., & Chien, F.-Y. (2001). The design of a lexical difficulty filter for language learning on the Internet. In T. Okamoto, R. Hartley, J. Kinshuk, & Klus, J. (eds.), *Advanced learning technology: Issues, achievements and challenges* (pp. 53–4). Los Alamitos, CA: IEEE Computer Society.
Lee, D. & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes* 25, 1, 56–75.
Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (eds.), *Teaching and language corpora* (pp. 1–23). London: Longman.
Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker et al. (eds.), *Text and technology* (pp. 157–76). Amsterdam/Philadelphia: John Benjamins.
Ma, B. (1994). Learning strategies in classroom concordancing. In L. Flowerdew & K. Tong (eds.), *Entering text* (pp. 197–226). Hong Kong: Hong Kong University of Science and Technology and Guangzhou Institute of Foreign Languages.
MacWhinney, B. & Snow, C. (1991). The *CHILDES project: Tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum.
Mauranen, A. (2006). English in the hands of non-natives: What's going on? Plenary address presented at ICAME 27, Helsinki, May.
McEnery, T. & Wilson, A. (1997). Teaching and language corpora. *ReCALL* 9, 1, 5–14.
McEnery, T. & Wilson, H. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
Milton, J. (1998). Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment. In S. Granger (ed.), *Learner English on computer* (pp. 186–98). London/New York: Addison Wesley Longman.
Milton, J. (2006). Resource-rich web-based feedback: Helping learners become independent writers. In K. Hyland & F. Hyland (eds.), *Feedback in second language writing: Contexts and issues* (pp. 123–39). Cambridge: Cambridge University Press.
Nattinger, J. (1980). A lexical phrase grammar for ESL. *TESOL Quarterly* 14, 337–44.

Nattinger, J. & de Carrico, J. (1992). *Lexical phrases and language teaching.* Oxford: Oxford University Press.
O'Keeffe, A. & Farr, F. (2003). Using language corpora in initial teacher education: Pedagogic issues and practical applications. *TESOL Quarterly* 37, 3, 389–418.
Pecorari, D. (2003). Good and original: Plagiarism and patchwriting in academic second-language writing. *Journal of Second Language Writing* 12, 317–45.
Pecorari, D. (2006). Visible and occluded citation features in postgraduate second-language writing. *English for Specific Purposes* 25, 4–29.
Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English Language*. London/New York: Longman.
Richards, J. C. & Rogers, T. (2001). *Approaches and methods in language teaching*. Cambridge: Cambridge University Press.
Rule, S. (2004). French interlanguage oral corpora: Recent developments. In F. Myles & R. Towell (eds.), *The acquisition of French as a second language.* Special issue of *Journal of French Language Studies* 14, 3, 343–56.
Rule, S., Marsden, E., Myles, F., & Mitchell, R. (2003). Constructing a database of French interlanguage oral corpora. In D. Archer, P. Rayson, E. Wilson, & T. McEnery (eds.), *Proceedings of the Corpus Linguistics 2003 Conference*, UCREL Technical Papers no. 16 (pp. 669–77). University of Lancaster.
Scott, M. & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language teaching*. Amsterdam/Philadelphia: John Benjamins.
Sealey, A. & Thompson, P. (2004). "What do you call the dull words?" Primary school children using corpus-based approaches to learn about language. *English in Education* 38, 1, 80–91.
Sinclair, J. M. (1990). *Collins COBUILD English grammar.* London: HarperCollins.
Sinclair, J. M. (1991). *Corpus, concordance collocation*. Oxford: Oxford University Press.
Sinclair, J. M. & Renouf, A. (1988). A lexical syllabus for language learning. In R. Carter & M. McCarthy (eds.), *Vocabulary and language teaching* (pp. 140–60). London: Longman.
Stenström, A., Andersen, G., & Hasund, K. (2002). *Trends in teenage talk: Corpus compilation, analysis and findings*. Amsterdam/Philadelphia: John Benjamins.
Stevens, B. (1991). Concordance-based vocabulary exercises: A viable alternative to gap-fillers. In T. Johns & P. King (eds.), Classroom concordancing. Special issue of *English Language Research Journal* 4, 47–61.
Stubbs, M. (1996). *Text and corpus analysis*. Oxford: Blackwell.
Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.
Stubbs, M. (2004). Language corpora. In A. Davies & C. Elder (eds.), *The handbook of applied linguistics* (pp. 106–32). Oxford: Blackwell.
Swales, J. M. (ed.) (1988). *Episodes in ESP*. London: Prentice-Hall.
Swales, J. M. (1990). *Genre analysis: English in academic and research settings.* Cambridge: Cambridge University Press.
Swales, J. M., Ahmad, U. K., Chang, Y.-Y., Chavez, D., Dressen, D. F., & Seymour, R. (1998). Consider this: The role of imperatives in scholarly writing. *Applied Linguistics* 19, 1, 97–121.
Thompson, G. (1993). *Reporting*. London: Collins.
Thurston, J. & Candlin, C. (1993). *Exploring academic English: A workbook for student essay writing*. Sydney: National Centre for English Language Teaching and Research.
Tribble, C. (2000). Genres, keywords, teaching: Towards a pedagogic account of the language of project proposals. In L. Burnard & T. McEnery (eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 75–90). Frankfurt am Main: Peter Lang.

Tribble, C. & Jones, G. (1990). *Concordances in the classroom*. London: Longman.
Tsui, A. (2004). ESL teachers' questions and corpus evidence. *International Journal of Corpus Linguistics* 10, 335–56.
Weber, J. J. (2001). A concordance- and genre-informed approach to ESP essay writing. *ELT Journal* 55, 1, 14–20.
Wible, D., Kuo, C.-H., Chien, F.-Y., Liu, A., & Tsao, N.-L. (2001). A web-based EFL writing environment: Integrating information for learners, teachers, and researchers. *Computers and Education* 37, 297–315.
Wichmann, A. (1995). Using concordances for the teaching of modern languages in higher education. *Language Learning Journal* 11, 61–3.
Widdowson, H. G. (2004). *Text, context, and pretext: Critical issues in discourse analysis*. Oxford: Blackwell.
Wilkins, D. A. (1972). *Notional syllabuses*. London: Longman.
Willis, D. (1990). *The lexical syllabus: A new approach to language teaching*. London: HarperCollins.
Willis, J. & Willis, D. (1989). *Collins COBUILD English course*. London: Collins.
Yoon, H. & Hirvela, A. (2004). ESL student attitudes towards corpus use in L2 writing. *Journal of Second Language Writing* 13, 257–83.

FURTHER READING

Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly* 34, 438–60.
Fox, G. (1998). Using corpus data in the classroom. In B. Tomlinson (ed.), *Materials development in language teaching* (pp. 25–43). Cambridge: Cambridge University Press.
Swales, J. M. (2002). Integrated and fragmented worlds: EAP materials and corpus linguistics. In J. Flowerdew (ed.), A*cademic discourse* (pp. 150–64). London: Longman.