

Corpora and Language Teaching: Just a fling or wedding bells?

March 2005 — Volume 8, Number 4

Corpora and Language Teaching: Just a fling or wedding bells?*

Costas Gabrielatos
<c.gabrielatos@lancaster.ac.uk>
Department of Linguistics and English Language,
Lancaster University, UK

Abstract

Electronic language corpora, and their attendant computer software, are proving increasingly influential in language teaching as sources of language descriptions and pedagogical materials. However, few teachers are clear about their nature or their relevance to language teaching. This paper defines corpora and their types, discusses their contribution to language learning and teaching, and provides examples of their use in class. It also outlines the changes in knowledge, skills and attitudes that are needed for learners and teachers to take advantage of the opportunities offered by the availability of corpus resources. Finally, the paper discusses the limitations of using corpora in language teaching, and the potential pitfalls arising from their uncritical use. Although the paper refers to research and teaching materials and procedures relevant to English language teaching (ELT) it addresses issues related to language teaching in general.

Introduction

Corpora first came to the attention of most English language teachers in 1987 with the publication of *Collins COBUILD English Language Dictionary*, the first corpus-based dictionary for learners. The following year saw the publication of an influential paper on the use of corpus-derived and corpus-based materials in the language classroom (Johns, 1988), although these had been proposed earlier (e.g., Higgins & Johns, 1984; Johns, 1986; Leech, 1986; McKay, 1980; Sinclair, 1986).

Since then, corpus-based language studies and pedagogical materials have grown exponentially; there is already a substantial and ever-growing body of corpus-based research on language structure and use, as well as on language learning and teaching (see Biber et al., 1998; Hunston, 2002; Kennedy, 1998; McEnery & Wilson, 2001; McEnery et al., 2005, in press; Meyer, 2002; Partington, 1996; Stubbs, 1996, 2001; Tognini-Bonelli, 2001). [1] 'Corpus' has now become one of the new language teaching catchphrases, and both teachers and learners alike are increasingly becoming consumers of corpus-based educational products, such as dictionaries and grammars. However, few teachers are clear about the nature of corpora, or their significance for language teaching, and fewer still have ever made direct use of a corpus. The questions most frequently asked by teachers are: *What is a corpus? How are corpora relevant to language teaching? How*

can they be used? The first aim of this paper is to answer those questions, provide an outline of the current state of affairs, and give examples of corpus types and uses. [-1-]

The utility of corpora for language teaching has been questioned from different perspectives. The sceptics have expressed reservations about the ability of corpora to capture language use (e.g., Widdowson, 1991), or the usefulness of native-speaker (L1) corpora in providing a model for teaching (e.g., Prodromou, 1997), some going so far as to argue that L1 corpora can intimidate learners (Gabbrielli, 1998), or disempower teachers (Dellar, 2003). Conversely, the fact that corpus-based studies relevant to language learning concentrate on those issues into which the use of corpora can offer insights may be misinterpreted as implying that corpora are the be all and end all of language teaching. [2] The second aim of this paper, therefore, is to demystify corpora and define their place within language teaching as a whole.

Corpus-based research and teaching have been carried out predominantly at universities; therefore, teachers in other educational settings may think that corpora are not relevant to their teaching situation, or that the knowledge, skills and technology required to integrate corpora into their teaching are beyond them. However, there have been articles on how teachers with minimal computer resources can make use of corpora (c.f. Johns, 1991a, 1991b; Stevens, 1995; Tribble, 1997a, 2000). The third aim of this paper, then, is to demonstrate that using corpora is not an either/or option, but that teachers in different contexts can make use of them to different degrees to suit their learners and facilities.

Corpora: Nature and types

What is a corpus?

Loosely defined, a corpus is “any body of text” (McEnery & Wilson, 2001, p. 197), that is, any collection of recorded instances of spoken or written language. For example, a pile of written assignments (e.g., essays) waiting to be marked is, roughly speaking, a corpus. Let us assume that these assignments have been written by students about to start a language course, and that the teacher has not taught the students before. The teacher can read the essays to form a general impression of the strengths and needs of the new class, but he/she may also want to focus on specific areas of interest. For example, while reading the assignments, the teacher may realise that the learners frequently make collocation errors. In order to examine the problem more closely, the teacher can go through the assignments, locate and list the unacceptable collocations, and determine whether there are any recurring patterns, that is, whether learners need help with the collocations of particular words, perhaps words normally associated with the topic of the assignment.

In the case of a single class of twenty learners, this analysis might be somewhat time consuming, but it would still be manageable. If, however, there were one hundred assignments, the task would become impractical. However, if the learners had submitted their assignments in electronic form, and if the relevant software were available, the teacher could examine the use of specific collocations in a hundred or more scripts in the same time it takes to manually examine twenty. Better still, the teacher could observe more complex and detailed patterns, and with greater accuracy. Moreover, this electronic corpus would be a helpful resource for the teacher, as it would be available in the future for the examination of other language aspects. The corpus could also grow by the addition of new assignments, in which case the teacher could trace the learners’ development in given areas. This is why ‘corpus’ is currently understood as “a body of machine-readable text” (McEnery & Wilson, 2001, p. 197). [-2-]

Imagine that at the end of the course our hypothetical teacher decides to summarise his/her findings on the learners’ use of collocations and present them at a conference or in an article. How helpful would the findings be to teachers in other contexts? In other words, how valid would it be to generalise from these findings? Such a presentation or article would be useful, but obviously any conclusions should be treated with caution, because the findings would only reflect the specific group of learners, taught by the specific teacher, in the specific geographical and social context. Also, the findings would reflect the use of collocation in the learners’ writing rather than their speech, and their use in *specific* text types.

If the corpus contained texts by learners from all over a particular region, then it would be possible to draw more reliable conclusions. Still, the corpus compilers would need to include texts written by learners of the same level, and ensure that the texts were of the same type and on the same topics. In other words, the corpus would need to be *representative* of the type of learners and texts that they wanted to examine (see Biber, 1993). Also, as it would not be feasible or practical to collect texts by all the learners of the same level in the region, the corpus compilers would have to select a *sample* of texts from each class.

The same principles apply to native-speaker corpora. ‘A corpus of English’ raises the question, ‘Which variety of English?’ Even if we restricted ourselves to one variety (e.g., American or British English), it would be impossible to create a corpus of the whole language, not

least because language evolves continuously. We can only collect a sample, and strive to make this sample as representative as possible. This leads us to the stricter and much more helpful definition of a corpus as “a finite collection of machine-readable texts, sampled to be maximally representative of a language or variety” (McEnery & Wilson, 2001, p. 197).

Types of corpora

Corpora come in many shapes and sizes, because they are built to serve different purposes. [3] There are two philosophies behind their design, leading to the distinction between *reference* and *monitor* corpora. Reference corpora have a fixed size; that is, they are not expandable (e.g., the British National Corpus), whereas monitor corpora are expandable; that is, texts are continuously being added (e.g., the Bank of English). Another design-related distinction is whether a corpus contains *whole texts*, or merely *samples* of a specified length. The latter option allows a greater variety of texts to be included in a corpus of a given size.

In terms of content, corpora can be either *general*, that is, attempt to reflect a specific language or variety in all its contexts of use (e.g., the American National Corpus), or *specialised*, that is, aim to focus on specific contexts and users (e.g., Michigan Corpus of Academic Spoken English), and they can contain *written* or *spoken* language. Corpora can also represent the different varieties of a single language. For example, the International Corpus of English (ICE) contains one-million-word corpora representative of different varieties of English (British, Indian, Singaporean, etc.). As implied in the previous section, corpora may contain language produced by *native* or *non-native speakers* (usually learners). Finally, corpora can be *monolingual* (i.e., contain samples of only one language), or *multilingual*. Multilingual corpora are of two types: they can contain the *same text-types* in different languages, or they can contain the *same texts translated* into different languages, in which case they are also known as *parallel corpora* (Hunston, 2002; Kennedy, 1998; McEnery & Wilson, 2001; Meyer, 2002). [-3-]

Creating a useful corpus

First, the texts a corpus is to contain are selected and stored in electronic format. Written texts, if they are not already in electronic form (e.g., downloaded from the Internet, submitted by learners on a disc or CD-ROM, or sent by e-mail), must be scanned; spoken texts must be recorded and transcribed. [4] The result of this stage is a *raw* corpus. Although a raw corpus can yield some information about language use, its usefulness is limited. For example, although the frequency of the word *drive* in the raw corpus can be determined, we will not know how many times it occurs as a noun and how many as a verb. Of course, different instances could be counted manually, but this would defy the purpose of compiling a corpus.

The utility and flexibility of a corpus can be increased by adding coding that a computer can recognise. Labels (or *tags*) are attached to the words, phrases, sentences, paragraphs, sections, or to entire texts in the corpus. Information related to non-linguistic properties of the texts is referred to as *mark-up*. Mark-up may give information about the source of the text (e.g., book, newspaper), the date of publication or broadcast, the author or participants, or text sections (e.g., introduction, conclusion). Information related to the linguistic properties of the texts in the corpus is called *annotation*. Most L1 corpora are annotated for the part of speech and form of the words (e.g., singular/plural, present/past tense). This type of annotation is also called grammatical annotation, or *tagging*. For example, the word *teaching* would be tagged ‘teaching_VVI’ if it was a present participle (as in ‘she was teaching’), and ‘teaching_NN1’ if it was used as a noun (as in ‘language teaching’). Corpora can also be annotated for lexical sense (e.g., lexis denoting belief, expectation) and pragmatic function (e.g., request, invitation). [5] What kind of mark-up or annotation is added to a corpus is determined by the information to be extracted. Sample 1 shows the three questions asked in the second paragraph of this article, annotated for part of speech. [6]

What_DDQ is_VBZ a_AT1 corpus_NN1 ?_?

How_RRQ are_VBR corpora_NN2 relevant_JJ to_II language_NN1
teaching_NN1 ?_?

How_RRQ can_VM they_PPHS2 be_VBI used_VVN ?_?

Sample 1. Example of annotation for parts of speech

How are corpora relevant to language teaching?

Corpus use contributes to language teaching in a number of ways (Aston, 2000; Leech, 1997; Nesselhauf, 2004). The insights derived from

native-speaker corpora contribute to a more accurate language description, which then feeds into the compilation of pedagogical grammars and dictionaries (Hunston & Francis, 1998, 1999; Kennedy, 1992; Meyer, 1991; Owen, 1993). The analysis of learner language provides insights into learner needs in different contexts, which then inform learner dictionaries and grammars. Research on learner corpora also contributes to our understanding of language learning processes (Granger et al., 2002). Corpora of language teaching coursebooks enable the examination of the language to which learners are exposed, and, when compared to L1 corpora, facilitate the development of more effective pedagogical materials. Learner corpora have the potential to contribute to the construction and evaluation of language tests in a multitude of ways (see Alderson, 1996); however, this potential has remained underexploited (but see Ball, 2001; Barker 2004). Finally, both native-speaker and learner corpora can themselves be used as learning/teaching materials (Aston, 1997; Aston et al., 2004; Johns, 1991a; Kettemann, 1995). Figure 1 summarises the interconnecting ways in which corpora are relevant to language teaching (adapted from McEnery & Gabrielatos, 2005, forthcoming). [-4-]

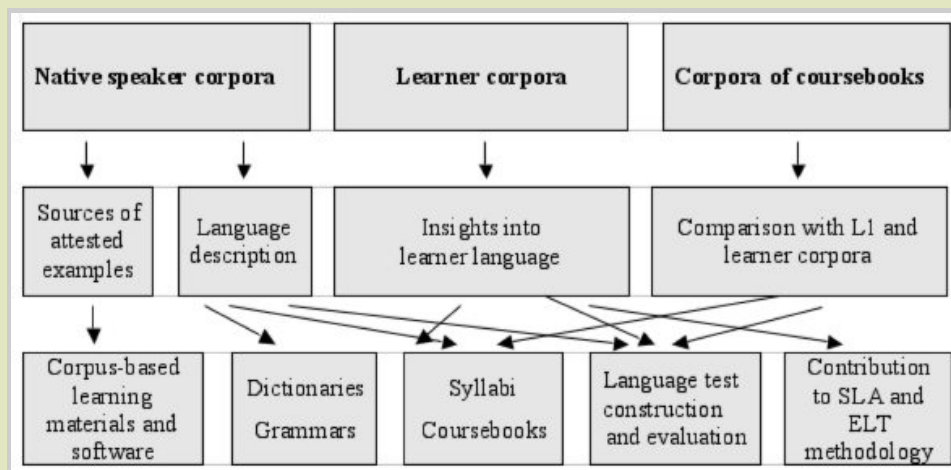


Figure 1. Corpora and ELT

We will now turn to the contribution of corpora to language teaching in more detail.

Language description

The use of L1 corpora in linguistic research has provided the most convincing evidence of discrepancies between actual use and traditional, introspection-based views on language (Sinclair, 1997, pp. 32-34), and has revealed patterns that had not been detected by introspection. This is pertinent to language teaching, as the information about language structure and use that learners receive, whether through pedagogical materials or teachers, is still largely based on introspection.

Helpful as it may be, introspection is not always reliable. Being a native speaker does not automatically mean that a user has a conscious, clear, and comprehensive picture of the language in all its contexts of use, nor do all native speakers share the exact same intuitions. A good example is the claim by a native-speaker teacher that in English, “question tags, along with bowler hats, mostly belong to 1960s BBC broadcasts” (Bradford, 2002, p. 13). This view is contradicted by the findings of Biber et al. (1999, p. 211), based on the examination of the 40-million-word Longman Spoken and Written English Corpus, who report that “about every fourth question in conversation is a question tag.”

It is, of course, very helpful to examine the intuitions of native speakers and elicit the different alternatives they find acceptable, or can generate by manipulating their language. It is equally helpful, however, to examine which of these alternatives native speakers actually use, and in what contexts and frequency. The discrepancy between intuitions and attested use indicates that when the language information learners are given is based only on intuitions, and when the examples and texts used in class are chosen to reflect these intuitions, then teachers and materials writers may unwittingly present their personal informal observations about language as the true and full picture of language structure and use, or present their own preferred usage as the only ‘correct’ or ‘acceptable’ one. The importance of corpus-informed pedagogical materials becomes more evident if we take into account that “to a great extent, the course-book can be considered to be the learners’ ‘corpus’” (Gabrielatos, 1994a, p. 14). [-5-]

Corpus-based research has also revealed the inadequacy of many of the rules that still dominate ELT materials. For example, in a study of a random sample of 710 *if*-conditionals [7] from the written section of the BNC, the conditional sentences were examined against the

information about form, time orientation and attitude to likelihood given within the currently favoured framework of five types (zero, first, second, third and mixed). The rules presented in fifteen recent intermediate-to-advanced coursebooks, taken collectively, accounted for only 44% of the sentences (Gabrielatos, 2003b). [8]

This section has highlighted the first important contribution of corpus-based research to language teaching, namely more accurate descriptions of English, which in turn can inform reference books and pedagogical materials (Hahn, 2000; Mindt, 1997). The language insights derived from corpora go beyond questions of correct or natural use, and provide additional details about the frequency of particular language features in specific contexts.

Examining learner language

Strange as it may sound, every single teacher has used a learner corpus, in the loose definition, if only in an informal and intuitive way. Teachers routinely write end-of-course reports, or answer questions about a learner's strengths and needs. How are they able to do so? To use corpus terminology, each learner's performance during the course is used to compile what we may call a *mental corpus*, which is consulted when evaluating a learner. The same applies when assigning an impression mark to a piece of writing or a task performance. Using language corpora allows teachers to be much more precise in examining learner language and identifying needs than just forming an overall impression, because corpus use enables teachers to examine particular areas in detail, or annotate for specific learner errors (Granger, 1999).

In general, studies on learner language focus on the over/under use of specific features in different contexts in comparison to native-speaker use, and the analysis and categorisation of learner errors. Error analysis may deal with frequent or common errors, or error patterns, according to the learners' L1, level and age, the medium of production (speech or writing), or the context of use (e.g., homework, test), while taking into account factors such as task and text type. Studies using learner corpora have focused on diverse aspects of learner language, mainly in writing. Examples of areas that have been examined with the help of language corpora are the use of lexical chunks (De Cock et al., 1998), collocations (Nesselhauf, 2005), complement clauses (Biber & Reppen, 1998), the progressive and questions (Virtanen, 1997, 1998), overstatement (Lorenz, 1998), connectors (Altenberg & Tapper, 1998), speech-like elements in writing (Granger & Rayson, 1998), and epistemic modality (McEnery & Kifle, 2002).

One area of language teaching which has interested corpus researchers is English for Specific/Special Purposes (ESP), especially English for Academic Purposes (EAP). [9] The areas that have most attracted corpus-based research are those of scientific and academic writing, often with a view to the implications for teaching (Coxhead, 2002; Flowerdew, 2002). In scientific/academic writing, the term 'learner' can be interpreted in two ways: a learner of the language system as a whole, or a learner of the style and conventions of academic writing. It is interesting that the latter applies to non-native speakers (NNS) and native speakers (NS) alike, in that both groups are, in several respects, approached as "trainee academics," the writing of which is "compared to that of established writers as evidenced in the discourse of published papers" (Gabrielatos & McEnery, 2005, in press, p. 312). The blurring of the NS-NNS distinction, as far as academic writing is concerned, is better understood if we consider that NNS who have published academic/scientific papers must be considered as "established writers" (Gabrielatos & McEnery, 2005, in press, p. 312; see also Lucas et al., 2003). Studies on academic and scientific writing have focused on language features, such as directives (Hyland, 2002), modality (Hyland & Milton, 1997; Thompson, 2002), or collocations (Gledhill, 2000; Luzon Marco, 2000), as well as the conventions of academic writing, such as citation practices (Harwood, 2004; Hyland, 1999; Thompson & Tribble, 2001). Finally, corpora can be used to detect plagiarism in student essays (Atwell et al., 2003; Lyon et al., 2004; van Halteren, 2003). [10] [-6-]

The contribution of such studies is two-fold. By examining learner language, we can define areas that need special attention in specific contexts and at different levels of competence, and so devise syllabi and materials. The analysis of learner language can also provide insights into the process of language learning (Bekiou & Diaz, 2004; Tono, 2000).

Corpora, language exposure, intuitions and generalisations

A corpus in the mind?

Intuition, or 'a feel for the language,' is what learners aim to develop. Native speakers develop that 'feel' partly through exposure to language in use and the recognition of patterns. Through this exposure, native speakers build the mental equivalent of a corpus (Bod, 1998). Intuitions can be seen as the results of the informal analysis of this *mental corpus*. It follows then, that by working on representative examples from language corpora, learners will be helped to recognise recurring patterns of structure and meaning. As Stern states, language learners need to be helped "to see a particular feature ... not merely as an isolated item but as part of an evolving system of

interrelationships which should become increasingly differentiated as it grows” (1992, p. 145). The wealth of instances of use of a specific item that corpora provide can offer the amount of evidence required for learners to refine their perception of it.

Pattern recognition, generalisations and rules

This section will first use a visual example to illustrate how pattern recognition works, and then discuss the implications for language teaching and the use of corpora, with particular regard to the formulation of pedagogical rules. We will assume that the images used in this example represent a specific language feature, such as the use of a grammatical structure, or the collocational behaviour of a word. We will also assume that we wish to establish the behaviour of the feature by examining a small number of language examples. On the strength of the analysis of this sample, we recognise a regular pattern (Figure 2).

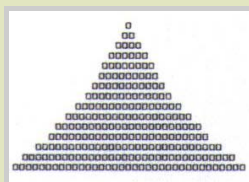


Figure 2.

In traditional language teaching fashion, we could formulate a rule. However, it might be that when more examples are added to the sample, some irregularities emerge (Figure 3a).

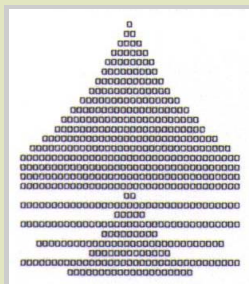


Figure 3a.

In the light of the new evidence, we could formulate a list of exceptions to our rule (Figure 3b).

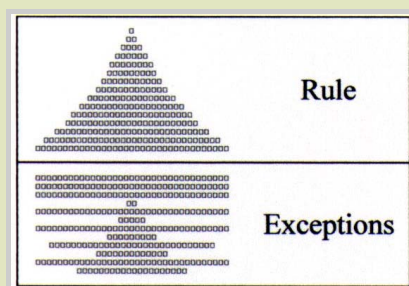


Figure 3b.

[-7-]

Let us assume that, over time, we come to observe more instances of the language item in question, or, in corpus terms, that we examine a larger sample, and that our observations reveal even more irregularities to the initial pattern, or, in language teaching terms, more exceptions to the rule (Figure 4).

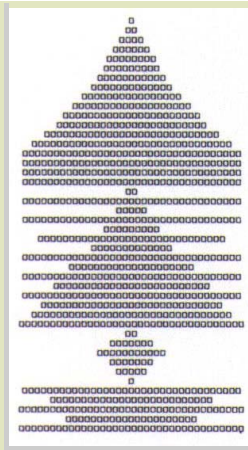


Figure 4.

On the face of the evidence at this point, two alternatives exist: First, we can conclude that the particular language feature is “illogical,” and that even if a rule could be formulated, it would inevitably have a disproportionate number of exceptions. Second, we could become suspicious of the fact that the exceptions cover more instances than the rule, and tentatively conclude that the fault lies with the rule, not the language. We could then hypothesise that what we have observed is only a part of a different pattern from that was initially perceived—a pattern that may be larger and more complex. If we adopt the second alternative, the next logical step is to further increase the size of the sample (Figure 5).

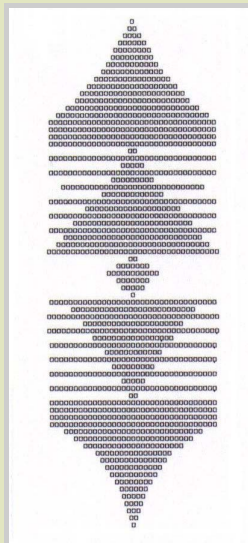
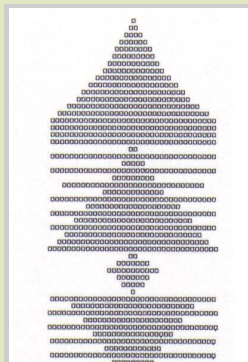


Figure 5.

The larger sample seems to reveal a new pattern. However, in the light of previous experience, this time we are not so quick to draw conclusions or formulate rules. Since the larger the sample, the more valid the conclusions, we considerably increase the sample size to test our new hypothesis (Figure 6). [-8-]



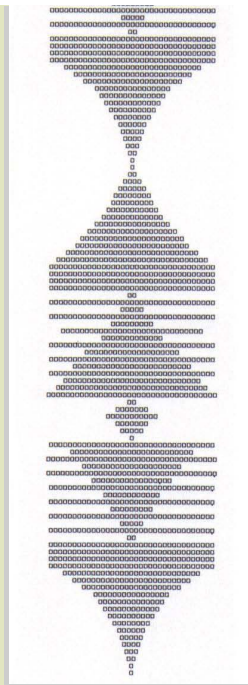


Figure 6.

Observing the pattern repeat itself (Figure 6), we are now in a much better position to formulate dependable generalisations about the language item in question. However, caution is needed regarding how these generalisations are delimited and phrased (c.f. Close, 1992, pp. 2-11; Leech, 1994; Swan, 1994; Westney, 1994). [11] The delimitation of generalisations relates to a number of important parameters that must be considered: [-9-]

1. The *medium*; that is, whether the sample contains only speech or only writing, or both.
2. The *context of use*, that is, “the physical, social and psychological background in which language is used” (Gabrielatos, 1999, p. 15). The main contextual elements are the topic, the writer’s or speaker’s purpose, the type of text or interaction, the audience or participants and their relationship.
3. The *co-text*, that is, the surrounding text or linguistic neighbourhood of the feature, as words and structures seem to both attract, and interact with, one another.
4. The *representativeness* of the sample; in other words, the collection of texts needs to represent a microcosm of the language use of the population under investigation.
5. The *size* of the sample; as the example demonstrated, language patterns may be too large and complex for a small sample to reveal adequately.

It would be rash to make broad statements about the behaviour of a language feature without reference to these parameters. As far as language teaching is concerned, exceptions and special cases are usually the result of overgeneralisations that do not take into account the parameters outlined above, or rules formulated on the basis of inadequate or selective evidence.

Corpora and condensed language exposure

Language learners in countries where the target language is not widely spoken often lack opportunities for the rich language exposure that is essential for developing the ability to recognise patterns. Extensive reading (Nation, 1997; Susser & Robb, 1990) is believed to facilitate language learning, because it exposes learners to real language use in context, and in amounts far larger than the short texts and dialogues usually preferred for the presentation of new language items. Extensive reading is also regarded as an effective way to help language learners develop intuitions as native speakers do (Krashen, 2004). The pattern-recognition example in the previous section gives an indication of how focused language exposure can be used actively, in order to formulate intuitions about language use.

Representative corpora can offer condensed exposure to language patterns. It is not argued here that corpora should be the sole vehicle for the development of reading skills and strategies, [12] nor is it argued that corpus use can replace out-of-class reading. Rather, what is being suggested is an approach that shares characteristics of both intensive and extensive reading—what might be called *condensed reading*. The reading of corpus samples is intensive in the sense that learners focus on the behaviour of specific language features; it is extensive in the

sense that learners examine language features in a larger number of texts than in conventional text-based techniques. Condensed reading enables learners to engage with language use in context in order to formulate and check, though not necessarily consciously, hypotheses about language structure and use.

One printed page contains 500 words on average. [13] The British National Corpus contains 90 million written words, or the equivalent of approximately 180,000 pages. A six-year language teaching programme of five one-hour lessons per week amounts to a total of about 1,000 lessons. To gain exposure through reading to the amount of language evidence contained in a 90 million word corpus, a learner would need to examine about 180 pages per lesson (in the case of classroom or intensive reading), or read about 80 pages every day of the year for six years (in the case of out-of-class or extensive reading), the equivalent of two to three books per week.

Through corpora, learners will experience types of texts that they may not choose to read out of class, or that teachers and materials writers may not deem appropriate. It seems clear, then, that learners may benefit from using corpora in addition to pedagogical materials and authentic texts. [14] The considerations listed here also highlight the limitations of pedagogies that avoid the use of materials and a pre-planned focus on language, such as the ELT translation of Dogme (Thornbury, 2000). These approaches tend to favour class discussions loosely structured around topics, with the teacher and learners acting as the main, or even sole, sources of language exposure. In doing so, they offer limited exposure to language, which is usually further restricted to the teacher's language variety and preferred usage. [-10-]

Corpora in the classroom

Before examining ways in which corpora can be used as (sources of) classroom materials, we need to clarify that a data-driven, awareness-raising approach is not necessarily linked to the use of corpora. Teachers can use texts containing the target language features and, through awareness-raising tasks, guide learners to discover the behaviour of lexical, grammatical or discourse elements. Therefore, it would be helpful to distinguish between *text-based* and *corpus-based* approaches to data-driven learning. [15]

Corpora can be used in language teaching in two ways (Leech, 1997, p. 10): The *soft version*, requires only the teacher to have access to, and the skills to use, a corpus and the relevant software. The teacher prints out examples from the corpus and devises the tasks. Learners work with these corpus-derived and corpus-based materials (Bernardini, 2004; Granger & Tribble, 1998; Osbourne, 2000; Tribble, 1997b; Tribble & Jones, 1990). Usually corpus examples are in the form of a *concordance*, where the word or structure being examined in the task is in the middle, so that patterns are more easily discernible (see Sample 2). The *hard version*, requires learners to have direct access to computer and corpus facilities and have the skills to use them (Aston, 1996). Tasks can be devised by the teacher (Tognini-Bonelli, 2001), contained within a CALL programme (Hughes, 1997; Milton, 1998), or chosen by the learners, with or without the teacher's guidance (Bernardini, 2002).

Taking into consideration the aims of a lesson, the design or selection of materials and the management of learning, in relation to teachers and learners, we can define combinations that cover the spectrum from totally teacher-centred to totally learner-centred. At the teacher-centred end, the teacher decides on the aims of the lesson, selects/designs the materials and manages the lesson. At the learner-centred end, the learner decides on all three, with the teacher or computer programme acting as facilitator and guide. Of course, there can be intermediate combinations, particularly when decisions are taken collaboratively between teacher and learners.

Soft version: Four examples

Example 1. Comparing text-based and corpus-based approaches to teaching collocations

This example shows how a text-based data-driven approach could be used to teach collocations of the noun *diet* to a group of intermediate-level learners. Because class time is limited, a long text or a small number of short texts could be used. Also, it would be wise to focus on a specific collocation pattern—only collocations of the noun *diet* in the singular with verbs, phrasal verbs, or expressions containing verbs, for example.

When selecting suitable texts, it becomes clear that it is difficult to find authentic texts which are 'about diet,' as they have not been written for language teaching purposes. The three texts chosen for this example [16] gave advice on dieting or reported on dieting experience. Although the texts are long for a typical 60/90-minute lesson (they total 2,250 words), they contained only 12 instances of the noun *diet*, and only 5 collocations with verbs, 2 of which were with the same phrasal verb (Sample 2). [17]

2	heart disease are from unhealthy	diet	- cardiac experts are keen to stre
3		diet	Guidelines Aimed at Healthy People
4	ent's suggestion they direct their	diet	advice to overweight Americans.
5	people, you begrudgingly go on a	" diet ."	Your initial concept of a
6	"Your initial concept of a	diet	is more commonly known as STARVATION
7	vicious cycle as you went from one	diet	to the next. Every new
8	Every new	diet	started with hope and promise, and
9	y thinking, "Oh great, another fad	diet	with a catchy name and empty promi
10	NO! The Eat and Burn	diet	identifies over 100 foods that tur
11	The Eat & Burn	diet	is easy to follow . You don't feel
12	concept behind The Eat and Burn	diet	is this: eat foods that safely forc

Sample 2. Concordance of the noun **diet** in the three texts

[-11-]

It does not seem worthwhile to spend the time it takes learners to read more than 2,000 words to teach only four collocations, particularly if it is uncertain whether these are among the most frequent ones.

This example illustrates a problem with text-based approaches: authentic texts do not conveniently contain enough instances of the patterns or structures on which teachers may want to focus. Additionally, a given text cannot be expected to necessarily contain the most frequent patterns, or to contain them in proportions that reflect their overall frequency in language use. Until recently, the only solution to this problem was to write texts specifically for pedagogical use, so that a sufficient number of instances of the target language features could be included. However, such texts tend to contain the target language features in unnatural proportions. It could be argued that if the frequency of the target features in pedagogical texts reflected actual use, that is, if the content of the texts was informed by corpus data, then these texts would be good teaching tools. Unfortunately, this is not the case. As the example above indicates, it is unlikely that every single text will reflect the overall frequency of a word, pattern or structure. Consequently, these putative corpus-informed pedagogical texts would be too densely, and so, unnaturally, packed with particular features. The process of incorporating an unnatural number of specific language items into the texts affects other elements of discourse. The result is a text that is as inauthentic as the traditional pedagogical texts and dialogues.

The same collocation pattern could be approached using a concordance from the BNC. One advantage of using a corpus is that the frequency of patterns can be expected to reflect real language use. Another benefit is that more detailed patterns can be investigated. For instance, learners may be presented with two sets of examples to examine: one with the pattern 'verb + preposition + any word + *diet*', and one with 'verb + article + *diet*' (Samples 3 and 4 respectively). [18]

try other foods, although I advise against a	diet	of all dried food.
might have had recently could be affected by your	diet	, or alcohol and cigarette consumption.
or whatever—of British women are on a	diet	at any one time, but that, as a nation
was breast-feeding, her doctor asked about her	diet	and found that she was a vegetarian. She
margarines, and are best avoided on this	diet	programme.
've been up when I've been on a	diet	, I mean smelling, smoked out, smoked out on

the end. When he had been on this	diet	for ten days he was tested with various foods.
her child demands breastfeeding despite being on a	diet	of solids. Concerns about dehydration if the child
feels it will be able to cook with a	diet	sweetener."
commune in Vancouver, Canada and fed on a	diet	of black pudding and Ecstasy. If I were to
legumes, meat hardly ever figuring in their	diet	. On as little as 8/6d (42 new pence
testing. He gradually forgets about the	diet	. This pitfall can be avoided by ensuring that
with the new knowledge I had gained about my	diet	I was eating sensibly, I no longer crave sweet foods
last saw you. You should go on a	diet	. Exercise more. Edouard rides every morning
the purpose of its use (to go on a	diet	, or to exclude certain elements such as meat)?
"I can go on a	diet	when I grow up," I said, but I was
scale is the seasonal dieter who goes on a	diet	in spring to get rid of the Christmas over-indulgence;
After going on the	diet	ask yourself these questions again—you may be
our current eating habits and including in our	diet	the necessary changes that are required to maintain a
You can help by getting involved in her	diet	, preparing healthy, balanced meals and emphasising that
unlikely that you are going to keep to the	diet	for very long. However, here is the vital
More people are killed by poor	diet	than by smoking, alcohol, drugs, accidents and
. The opposition can no longer live on a	diet	of anti-Thatcherism. They face a prime minister,
daren't eat chewing gum if I'm on a	diet	Oh you don't need to
be suspect and was temporarily omitted from the	diet	. The patient returned to eating only foods that
for salt becomes less as you progress through the	diet	programme.
. Arthritic working-class guys raised on a	diet	of fish and chips and fags; they died of
If you're on a	diet	and you've found that you've hit problems, ring
She's erm, she's on a	diet	. Oh really? She's lost
That's why she's on a	diet	! Cos she doesn't
enjoyed it. It didn't seem like a	diet	—in fact, if there was one sentence that
enough how important it is to set aside from your	diet	foods you suspect or know cause you problems.
The more you are able to stick to a	diet	of natural foods—fruit and vegetables (raw if
will have heard that if you do stick _VVI to a	diet	and lose weight, then your metabolism will drop so
was losing her will power to stick _VVI to her	diet	. Anne had already trimmed down to a reasonable
metabolic rate. Providing you stick _VVB to your	diet	, and don't consume lost of extra calories,
be so easy that you can stick with the	diet	until all the weight is off.
Stage II must immediately be struck out of your	diet	. This is very important. Failure to

to the calorie-counting method, supervised by a	diet	club, dietitian, or doctor.
, it dwelt in woods, surviving on a	diet	of maize, fruit and grass. The north American
experiment in which a doctor switched to a	diet	including the average adult consumption of the country's
it's fat, and you should think about a	diet	. But don't be bullied by precise,
like your neighbour's before he went on that	diet	.

Sample 3. verb + preposition + any word + **diet** (concordance view)

[-12-]

1. Therefore, **adapt** the **diet** according to your lifestyle, to your personal diet.
2. **Altering** the **diet** is also far more risky for a child than it is for an adult, so there are more difficult decisions to be made before embarking on an elimination diet.
3. Wholegrain cereals are also good for the same reasons, and protein foods(meat, fish, eggs and cheese) taken in moderation help to **balance** the **diet** and give all the necessary nutrients.
4. Even before you **begin** the **diet** , notice how fast you eat, and slow down.
5. You need to consider those antecedent events that prompt you to **break** a **diet** , and then think about which of these things you can avoid or change in some way.
6. The final events that lead our dieter to **break** the **diet** are quite concrete, namely, walking into a café, seeing and smelling the pastries, and seeing other people happily enjoying them.
7. I could manage to lose five or six pounds and then I would **break** the **diet** , go back to normal food and put all the weight back on again.
8. Unfortunately the punishment for **breaking** a **diet** is also in-built; you put weight back on.
9. What if I get a reaction to a particular food after I have **completed** the **diet** ?
10. Treatment was not merely a matter of prescribing herbal medicines, but a whole regimen which **controlled** the **diet** and the life-style.
11. Maybe it is due to my always having **eaten** a **diet** rich in red meat and saturated animal fats?
12. In addition, oily fish is a rich source of the Omega-3 fatty acids and recent medical research suggests that there are a number of health benefits to be gained from **eating** a **diet** rich in these fatty acids.
13. During the final two-week period you will be **eating** a **diet** composed of the foods you have selected through trial and error in the preceding four weeks.
14. Once you have **established** a **diet** on which the child remains well, be careful not to allow too much of any one food.
15. But, as RICHARD BATH discovers, England's appointment of coach DICK BEST for less than a year means that, instead of a bright new era, we can **expect** a **diet** of pragmatism and playing the percentages.
16. For example, it can be argued the expansion into amalgamated police units has enlarged the organization to a point where it is no longer accessible to the man in the street; alternatively, it may be that the use of a centralized computer and complex technical aids has alienated the public even at the same time they are increasingly **fed** a **diet** of violent news snippets which reinforce a fear of crime and generate another "folk devil" of criminal menace, which demands the impossible: a policeman on every corner.
17. In contrast, those abroad, notably in the West, who were **fed** the **diet** of stage-managed events, found his assassination both momentous and incomprehensible.
18. **Feed** a **diet** of insects, worms, plant matter, flake food and freeze dried food.
19. What we are going to do is **find** a **diet** that not only helps you to achieve effective weight loss, but is really healthy, suits your individual needs, and can be followed for years to come in order to maintain the weight and shape you want.
20. Therefore when you **finish** a **diet** , returning to eating an average amount, the amount you used to eat to maintain a constant weight, will result in you putting on fat.
21. But I doubt very much whether there are any claims now outstanding which are not statute-barred, in respect of children stillborn before 22 July 1976 or any children born before that date, who are locked in litigation with their mothers over whether the mother tasted alcohol or **followed** a **diet** other than that recommended by the current phase of medical opinion during pregnancy.
22. Nevertheless, I now had 120 people, 116 women and 4 men, who had **followed** the **diet** for a full eight-week period.
23. I could fill a book with the other similar comments which were written on the questionnaires but I think we can take it as read that the trials proved beyond doubt that if you **followed** the **diet** moderately strictly you could definitely lose inches from parts usually untouched by normal dieting methods.
24. You will essentially be **following** the **diet** in Stage I, but adding to it any food or drink you now have listed in column 3 of the Grand Review Chart on page 228.
25. So I thought, "I'll **invent** a **diet** where you feel good and you can eat."
26. You may prefer to **keep** the **diet** simple during your working week and to save the more elaborate meal for the weekends of, if you feel really adventurous, for dinner parties.
27. Unless you absolutely hate cooking, or are just too busy, it is preferable that you experiment with some of the recipes in order to **keep** the **diet** interesting.
28. Unless you absolutely hate cooking it is advisable to experiment with some of the recipes in order to **keep** the **diet** interesting.
29. The answer, therefore, is to **maintain** a **diet** which is as balanced and healthy as possible.

30. Ethnic minorities will have the right to **obtain** the **diet** required by their religious beliefs.
31. I remain nervously aboard, to hear doctors exchanging advice on every deck: Doctor McRae **recommends** a **diet** of rice and yoghurt.
32. For example, the origin of ivory can be identified by its strontium isotopic composition, which **reflects** the **diet** of the elephant.
33. In no circumstances should you do this without help and advice from your doctor—**restricting** the **diet** of small children can be very dangerous.
34. Some years ago Maisie had swallowed a whole bottle of vitamin pills and, although Henry had suggested that in his view Maisie’s stomach could probably have **stood** a **diet** of broken glass, aspirin and raw steak, Elinor had insisted on ringing Charing Cross Hospital.
35. Just having a little treat here or there can add up to enough to **stop** the **diet** from working.
36. The choice of diet—Two dietary factors—**supplementing** the **diet** with wheat bran and reducing the intake of refined carbohydrates—reduce cholesterol saturation of bile in subjects with supersaturated bile.
37. Robin-Anne nodded, but was too busy eating to take much notice of her brother, though she did manage to mumble that she **thought** the **diet** soda was really kind of good.
38. A software engineer who freely admits to being plump and has **tried** every **diet** in the book claims to have invented a programme that’s guaranteed to keep those extra pounds off.
39. Some of Dr Gerson’s patients—including those with TB—**tried** the **diet**.
40. My exercise class students witnessed the low fat diet’s remarkable effect on my body (I had lost only 6lbs [2.7kg] but all from my problem areas) and then they **tried** the **diet** with similar benefits.
41. Gerald was asked to **try** a **diet** containing no sugar or white flour and was given an anti-fungal drug, Nystatin.
42. In his forties it grew worse and he decided to see a specialist When Alan mentioned that he had taken a lot of antibiotics just before the urticaria began, the specialist suggested that he **try** a **diet** with no sugar and very little starch.
43. You are welcome to **vary** the **diet**, but do make sure you eat other foods besides chocolate this week!

Sample 4. VERB + ARTICLE + *diet* (sentence view)

[-13-]

Another important benefit of corpus use becomes apparent if we compare the number of different collocational patterns contained in the texts and corpus samples [19] in the example (Table 2).

	No. of words	No. of patterns
Texts	<i>2,250</i>	<i>5</i>
Corpus samples	<i>2,000</i>	<i>59</i>

Table 2. Comparison between texts and concordances

Although the texts and corpus samples have roughly the same number of words, the corpus samples contain twelve times the number of patterns. As mentioned above, pedagogical texts tend to contain an unnatural density of the target language features. The use of corpus samples achieves the same density, but without compromising natural use. The richness of the corpus samples makes it possible to devise tasks that cover a wide range of features. For example learners can be given the following task:

- Find verbs, phrasal verbs, or expressions containing verbs that combine with the noun *diet*.
- Does *diet* have the same meaning in every sentence?
- Which combinations seem to be more frequent with each meaning?
- Find combinations with similar/opposite meaning.
- Try to group the patterns in some way.

This task focuses on collocation patterns, lexical meaning and frequency of occurrence. It also involves some form of practice, or “mental contextualisation” (McCarthy, 1990, p. 36), as learners are asked to group the patterns in a meaningful way.

Example 2. Lexical inference

Corpus samples also lend themselves to work on reading skills, and, in particular, to developing strategies for inferring the meaning of unknown lexis in the text. Although it is, of course, possible to use one or more texts to train learners in this enabling skill, corpus samples are superior in a number of ways. A text will contain only a few instances of the lexical item, will usually demonstrate its meaning and use in one context, and may not provide sufficient clues for inferring meaning. Corpus samples, on the other hand, contain a large number of examples which demonstrate meaning and use in diverse contexts and offer a wealth of clues. Consider the following sample task (based on Sample 5 below):

In the following examples the same word is missing in each case.

- Work out the meaning of the missing word.
- Decide if the word has exactly the same meaning in all the sentences.
- If there are different meanings, check whether, and how, they are related.
- If there are different meanings, check whether there are specific words or expressions that are associated with each meaning. [-14-]

1. The winners of Black & Decker 9032 cordless ???????? action drills which were the prizes in a competition which appeared in the May issue of DIY are as follows:
2. It was a "maniacal" beating around the head with a claw ????????.
3. Endill watched Tock make a hole in the wall, holding his ???????? with both hands to stop it banging in the wrong place.
4. He had a ???????? and banged it against the walls to restore order but nobody took any notice of him.
5. It was taken out of the context of the early punks and placed alongside the ???????? and sickle, the IRA and PLO slogans and any other symbols which could be guaranteed to raise the hackles and the eyebrows of the BOF's (remember them?).
6. The problem could just be confined to this guitar and removing the strings and lightly tapping the frets down with a block of wood and a small ?????? ?? could well fix it.
7. Author Ian Fleming's original unpublished notes on his most famous creation are to go under the ???????? at London auctioneers Sotheby's on December 15.
8. There are several ways in which you can do this—I use a professional glazier's staple gun which is both quick and efficient, but if you find this too expensive an investment when you first begin pressed flower work you can use a ???????? and nails instead.
9. These mechanisms enable a stressed metal to be rapidly filled with dislocations (something like 10 per square centimetre) and thus to flow under a steady load or the blow of a ???????? quite easily.
10. Next Monday, Rod's prize goes under the ???????? through ADT, the world's biggest car-auction firm at Blackbushe, Hants.
11. Family's ???????? revenge on love cheat soccer ace
12. Chief union negotiator John Allen said it was another "???????? blow" for the industry.
13. But I must have felt the need for some support, because I found I'd grabbed hold of one of my ???????? s—a geologist is always armed with a ?????? ??—and when I got through to the back of the house he was there already, at the kitchen window."
14. Its plastic jacket bore a gold ???????? and sickle.
15. Jaq rifled through the pack to find the card he used to signify himself; the black-robed High Priest, enthroned, gesturing with a ????????.
16. The appellant, having discovered that the man had a number of previous convictions for similar offences, equipped himself with a ???????? and a quantity of weak sulphuric acid and sought out the man at his place of work on two occasions.
17. Using a small ???????? the needles were then tapped through these holes and then cut off flush with the exterior of the pipe.
18. "Distractedly, he began to change ????????, pouch, pipe and matches from hand to hand, dropping them and picking them up, before finally deciding to put the ???????? down and stuff the rest into his pockets.
19. She had laid the ???????? there, after she had tried to break the tower window.
20. The fist techniques of taekwondo involve lunge punches, reverse punches, back fists and ???????? fists—all of them similar to the basic karate punches described in the previous chapter.

Sample 5. Lexical inference

[-15-]

Example 3. Revision and critical examination of grammar rules

Corpus samples can be used for revision, and offer an opportunity for learners to formulate a second opinion on traditional ELT rules (see Leech, 1994). The following task focuses on *if*-conditionals and could be used with upper-intermediate and advanced learners. Only a small, random corpus sample is given here, which is too small to be representative. However, even in such a small sample, the limitations of the five-types framework become clear.

- Check if all *if*-sentences are conditional sentences.

- Do all conditional sentences express a cause-result relationship?
- Allocate each conditional sentence to one of the types in your course/grammar book. Do all sentences conform to the information/rules in your book?
- In your groups, discuss any sentences that don't fit the rules and report to the class.
- Try to adapt the rules so that they fit more sentences, or decide on a different way to group the sentences.

1. "My dear, dear fellow, if I had a lira for every time I've heard that story ... well ..."
2. If meat was banned, for instance, this was because the animal too has a soul (and may even be a dear departed relative!).
3. If Gunnell herself has cashed in, she's not been so blatant or obviously motivated by the financial side as other athletes.
4. If ordinary children build their linguistic abilities on antecedent social and cognitive abilities, these may, in fact, be necessary prerequisites for the emergence of language.
5. Payment of fines; imprisonment; amputation of right hand (the left hand is only amputated if the right has already been amputated)
6. The cold did little to hinder the Orcs, for Orcs and Goblins are hardy creatures, and, if needs must, will eat any flesh no matter how foul or what manner of creature it comes from.
7. If the lack of energy is not remedied, the excess stress on the body can ultimately lead to prolonged illness and possible death.
8. If you pull it off, I get fifteen hundred.
9. We shall examine random additions to a file; as the principles involved do not change if the additions are grouped or regular in pattern, the methods used can be adapted to suit those cases.
10. Into the power-vacuum created by the slaying of Osric and Eanfrith stepped Eanfrith's brother, Oswald, who slew Cadwallon in the battle of "Heavenfield" near Hexham in the autumn of either 634 (if Eadwine was killed in 633) or 635 (if Eadwine did not perish until 634), and assumed the kingship of both the Deirans and the Bernicians.
11. This enforced poverty made them easier targets for propaganda: if they left with no more than their allowance, they could be portrayed as shabby Untermenschen scuttling away like rats; if they managed to outwit the system, then they were economic criminals fleeing with stolen goods.
12. A. If you are a tenant of a public landlord, such as a local authority, new town or housing association, and if you have a pressing need to move to another local authority area for a job or social reasons (for example, because you are elderly or handicapped), you should ask your landlord whether you can be nominated for a move under the National Mobility Scheme.
13. "The facts speak for themselves; if Dana had any feelings for you she'd have refused my offer.
14. If a factory chimney dumps smoke on a thousand gardens nearby it may be very expensive to collect 1 from each household to bribe the factory to cut back to the socially efficient amount.
15. If the sale is by sample as well as by description it is not sufficient that the bulk of the goods corresponds with the sample if the goods do not also correspond with the description.
16. They are people whom we rarely consider in this House, but when there is a suicide or accident on the railway, the driver, and his mate if appropriate, may be mentally scarred for life by the experience.
17. The Member of Parliament shall be eligible for nomination for selection as the prospective parliamentary candidate and, whether nominated or not, he or she shall be entitled to appear as if they had been nominated before the special meeting of the General Committee convened in accordance with section (3) of this clause and to be considered for selection as the prospective parliamentary candidate.
18. Example 4:10 Tenant's power to make time of the essence (1) if the landlord fails to take any step in the procedure for rent review within a period of time prescribed by this lease (whether or not that step could also have been taken by the tenant) the tenant may give the landlord written notice.

19. It is perhaps as well to remember at the outset that the main injury in this particular case was a hip injury which, if it had occurred to a younger man, would have produced an arthrodesis operation.

20. "When I saw Ivo with a parcel he was about to mail to his wife's cousin in Karlovy Vary, I told him I was driving that way, and that I'd drop it into the shop where Edita's cousin works if he wished."

Sample 6. *If*-sentences (random sample)

[-16-]

Example 4. Homework tasks with a multiple focus

The variety of information in corpus samples can provide material for homework assignments. Learners can do the tasks outside of class so that classroom time can be devoted to feedback discussion and perhaps some fine-tuning by the teacher. For example, learners can be given corpus sentences with the words *sorrow* and *grief* (samples 7 and 8) [20] with a task that focuses on nuances of meaning, sense relations (synonymy and antonymy) and collocation patterns:

Examine the sentences with *sorrow* and *grief*.

- What causes *sorrow* or *grief*?
- What other words/expressions with a similar or opposite meaning can you find?
- What verbs, adjectives and nouns seem to combine more often with the two words?
- Are there any frequent fixed expressions?
- Devise a dictionary entry for each word.
- Look up the words in dictionaries and compare your entry with the information they give.

1. Fly over the solitary rock washed by the glacial tears of **sorrow**, let there be at your passing, a radiant beam over the gloomy solitary rock.
2. Among his sacred possessions were an enormous club which could raise the slain to life again; a magic harp whose music made its listeners forget **sorrow**; an inexhaustible cauldron from which no-one is turned away hungry; and two marvellous sheep – one eternally roasting, the other forever feeding in readiness for slaughter.
3. God promises his people comfort and invites us not to live in **sorrow**.
4. Dick expressed his great **sorrow** at the news of LF363 and said that he had a very soft spot for it, having "cut his teeth" on that Hurricane during his days with BBMF.
5. Another couple passed by giving them a wide and sympathetic berth, leaving them alone in their **sorrow**.
6. GE NUP DIMU The **sorrow** that overtakes the child in the womb when it knows it will be born dead
7. As I walked down by the riverside one evening in the spring Heard a long gone song from days gone by Blown in on the great North wind Though there is no lonesome corncrake's cry of **sorrow** and delight You can hear the cars and the shouts from bars and the laughter, and the fights May the ghosts that howled round the house at night never keep you from your sleep May they all sleep tight down in Hell tonight or wherever they may be ...
8. I do not need your understanding, or your damned **sorrow**!
9. "Through the night of doubt and **sorrow**, Onward goes the pilgrim band, Singing songs of expectation, Marching to the Promised Land."
10. The letters were handwritten in strong block capitals, and he peered at the sign for obvious traces of **sorrow**, such as shakiness of the hand or tear stains, but there were none.

Sample 7. *Sorrow* (random sample)

1. They hold their feelings in, their **grief**, anger, frustration and would rarely weep in front of others.
2. If, of course, her **grief** becomes prolonged and it seems that she is making no headway towards adjustment, it will be advisable for her to see her doctor, but this rarely happens.
3. There was movement that year among all the teams and Pace's death in March not only upset Brabham plans but also caused Ecclestone, who was very attached to Carlos, both as a man and as a driver, considerable **grief**.
4. Short-sighted Mansell, 26, came to **grief** on the hard shoulder of the M6 near Pontefract, West Yorks.
5. The second type had no power-had numbers, but no safety: numbers conferred only **grief** and weakness.
6. Mom gave birth to a baby boy, and called him Winchell (not after the doughnuts, though the association would later cause him **grief** at school).
7. We pass, heads to the side, in deep **grief** for a piece of shredded lorry tyre.

8. He also added that Moore had never asked them to forgive her for throwing their lives into **grief** and chaos.
9. I was distracted with **grief** this time, torn by guilt, and Eric had to look after me while I acted my part to perfection, though I say it myself.
10. In her despair and **grief** Mrs McDermott also turned to alcohol for relief.

Sample 8. *Grief* (random sample)

[-17-]

Soft version: Teacher manipulation of corpus examples

When using the soft version, teachers can manipulate the corpus examples in a number of ways. They can restrict the examples to a specific medium (writing/speech), and genre or text type (newspaper article, novel). They can also decide on the amount of text to give learners—only a few words on either side of the key word (as in Samples 2-3 above), an entire sentence (as in Samples 4-8 above), or a paragraph. Finally, they can edit the samples to remove sentences that they deem too difficult for the learners (Wible et al., 2002). This manipulation should be carried out with the understanding that the adapted samples are not good guides to the frequency of a language item.

We will take the pattern ‘verb + article + *diet*’ as an example. There are 3,458 instances of *diet* as a noun in the BNC, and 177 instances of the pattern. Since there would probably not be enough classroom time for learners to examine so many examples, the sample must be reduced (Sample 4 above contains 43 examples). If, for the sake of convenience, the first 43 examples from the original 177 were selected, the sample would not be representative, since the collocates are in alphabetical order. Instead, the collocations were extracted from a random sample of 1,000 sentences out of the total of 3,458. In this way, the 43 examples in Sample 4 give a much more accurate picture of the pattern. (However, the sample is too small to give a truly representative picture.) Also, in order to save time and keep a clear focus, instances of *diet* with the meaning ‘assembly’ were removed.

There are, of course, cases in which it is difficult to restrict the number of examples without affecting the representativeness of the sample, for example, when selecting examples for students of below-intermediate level. In this case, the teacher has three options: simplify the examples, select suitable sentences in a way that their make-up approximates the original sample, or avoid dealing with issues of frequency. Nevertheless, the edited sample may still be expected to contain at least some of the most frequent collocations.

Hard version: Learners using corpora

When learners have direct access to corpora, the focus of the lesson can be made more flexible to reflect their interests and needs. In other words, the teacher or learners have the option of modifying the aims and direction of the lesson on the spot according to what emerges. In the case of the collocations of *diet*, learners could also choose to examine other patterns, for example collocations of the noun *diet* with adjectives, patterns of *dieting* used as a noun, or *diet* as a verb. If the concordance or sentences do not offer enough clues, learners can get more text just by clicking either on the key word or a special button (depending on the software).

Corpora and ELT methodology

Although the use of corpora in language teaching has been linked to a “data-driven” approach (Johns, 1991a), it would be a mistake to assume that corpus use is restricted to any single teaching methodology. The use of corpora, in both the soft and hard versions, and either in a classroom context or for self-study, is compatible with all methodologies that accept explicit focus on language structure and use; in other words, teaching frameworks that reserve a role for noticing or awareness/consciousness-raising (e.g. Lightbown, 1985; Schmidt, 1990; Sharwood Smith, 1981). [-18-]

Corpus examples can enhance frameworks involving explicit presentation of language features, but they are particularly relevant to frameworks which depend on the learners using their existing language knowledge to work out the meaning and use of new elements (Rutherford & Sharwood-Smith, 1988), as has been shown by a number of studies utilising corpora as sources of language data (Aston, 1997; Granger & Tribble, 1998; Johns, 1997). Although it may not be readily apparent, corpus use is also compatible with methodologies that advocate exposure to language, or comprehensible input (Krashen, 1985), rather than explicit focus on language, as was demonstrated above through the example of condensed reading.

In other words, corpus use fits equally well within language-based approaches, with the Presentation-Practice-Production (PPP) framework as their best known realisation (Read, 1985; Spratt, 1985), and task-based approaches (Fotos & Ellis, 1991; Loschky & Bley-Vroman, 1993; Nunan, 1989; Skehan, 1998). In the case of a straightforward PPP lesson, corpus data can be used instead of made-up examples in the Presentation stage. But a corpus can only be utilised to the fullest if the PPP teaching framework has been modified and expanded to incorporate awareness-raising (Gabrielatos, 1994b) or data-driven procedures (Johns, 1997). Johns proposes a flexible sequence of *Research*, *Practice* and *Improvisation*, as he sees the learner “as ‘linguistic researcher’, testing and revising hypotheses, or as ‘language detective’, learning to recognize and interpret clues from context” (1997, p. 101). In formatted task-based frameworks, corpus data can be used in the “Pre-emptive Work” stage (Skehan, 1993), or the “Pre-task” and “Post-task” phases (Willis, 1996), which involve input or consciousness-raising.

Corpus use can also enhance learner independence. According to Johns (1997, p. 101), when using corpora or corpus-based materials, “students define their own tasks as they start noticing features of the data for themselves—at times features that had not previously been noticed by the teacher” (see also Bernardini, 2002). Along the same lines, the use of corpora enhances the use of the language lab, and suggests a more flexible and learner-centred use for CALL materials (McEnery et al., 1997). This is not to say that the teacher’s role is diminished; rather, it is enriched and diversified. The teacher becomes less a provider of input and facts about language and more a facilitator and consultant, or, at the learner-centred end, a co-researcher.

Finally, having learners work with samples from representative corpora of different varieties (e.g., British or American English) and different genres (e.g., academic English, chatroom English) will give them the rich exposure they need to become aware of the existence of varieties, not so much in order to learn these varieties, but to understand that English is not monolithic.

Corpus use in learning and teaching: Prerequisites

The availability of corpora and corpus software alone cannot ensure that language teaching will take full advantage of the opportunities they offer. Language teaching institutions will have to take certain courses of action; learners and teachers in their turn will have to adjust to changes in knowledge, skills and roles. [-19-]

What is apparent is the necessity for investment in computers, access to corpora, and the relevant software. This would be a costly move if a school were to opt for the hard version, but the cost would be reduced considerably if the soft version were adopted. In the first case there should be enough computers for each learner in a group, or at least for every two to three learners. In the second case, a school will only need enough computers for the teaching staff. Investment in technology, however, is just the tip of the iceberg; it is the investment in the *users* of corpora, the learners and teachers, that poses the greatest challenge for language teaching (see Kennedy & Miceli, 2001).

Learners need to become familiar with corpora (Leech, 1997, p. 10), and in the case of the hard version, they have to be trained to use corpus software (Bernardini, 2002). They also have to be introduced to data-driven approaches to learning, and guided to develop the skills that such approaches require. They have to be guided away from the “single correct answer” concept, and the notion of fixed rules and exceptions, towards the recognition of patterns and alternatives, and the importance of context. The utility of corpus use does not stop at helping learners discover language facts for themselves—when learners (are guided to) examine corpus samples they also develop a crucial element of learning skills (see Cohen, 2003; Oxford, 1994), namely the ability to recognise patterns of language structure and use. To employ a popular analogy, in consulting a dictionary or grammar learners are given fish; by actively engaging in pattern recognition they learn how to fish.

Of course, teachers need to be informed about corpora and the relevant software, and become skilled users (Renouf, 1997). This is not expected to take place quickly, and may be met with reluctance, or even resistance, on the part of teachers (Arkin, 2003). Teachers also need to be in a position to assist and guide learners in their language investigations. This means that the teachers’ awareness and knowledge of language will have to extend beyond the information in pedagogical materials (see Gabrielatos, 2002a, 2002b; Leech, 1994). Teacher preparation programmes would not only have to add components related to corpora and their uses, but also to place much greater emphasis on language awareness and description (see Andrews, 1994; Sinclair, 1982).

Maintaining a sense of perspective

English language teaching is vulnerable to pendulum swings, and has a propensity for the marketing and uncritical acceptance of “miracle methods” (see Decoo, 2001; Gabrielatos, 2001, 2003a). As mentioned in the introduction, many language teachers have little awareness of

issues pertaining to corpora and the analysis of naturally occurring data, and minimal, if any, familiarity with corpus software tools. [-20-]

Corpus evidence has challenged the over-reliance on intuitions that characterises much of language teaching. Shifting the focus on actual language use is clearly a positive development. However, it is conceivable that the language teaching pendulum may swing to the other extreme: an over-reliance on corpus data. Such *corpus worship* could lead teachers and learners to disregard the fact that, as large as corpora may be or may become in the future, they cannot capture the entirety of language use because, by definition, they are only samples (Gavioli, 1997, p. 85). It is also worth considering that corpus studies depend on labelling and counting language elements or learner errors, and that these labels themselves are informed by intuitions and linguistic theories (Sinclair, 2004). A more sensible attitude towards intuitions and corpora as sources of language insights would be neither that intuitions are useless, nor that corpora are the ultimate solution. As Sinclair (1991, p. 39) states, native-speaker introspections are useful “in evaluating evidence rather than creating it.” Therefore, intuitions should be balanced against, and enriched by, the evidence of language in use that corpora provide (see McEnery & Wilson, 2001, pp. 5-12). [21]

Corpus use, particularly in the form of concordances, is very well suited to the teaching of lexis and, to a lesser extent, grammar. As corpora and relevant software become more available, and corpus use becomes more widespread, language teaching may well concentrate on lexical and grammatical patterns, at the expense of discourse and interaction skills, that is, the teaching of reading, listening, writing, and speaking skills and strategies. Similarly, since corpus-based or corpus-derived materials are good vehicles for raising awareness of language features, language production and interaction may be given less than adequate attention. Differently stated, when working with corpora, learners become observers of language use. This is necessary for language learning, but not sufficient; learners also need to become participants in language use.

Corpus samples, and in particular random ones, may not be suitable to all teaching contexts. Consequently, teachers and materials/software developers may need to manipulate the samples given to learners, a process not without pitfalls, particularly when the focus is on the frequency of language features. For example, corpus samples may have to be adapted when used with low levels or young learners, and when corpus examples contravene or offend sociocultural norms and customs.

Corpora are also excellent sources of information about the frequency of language features in different contexts. Although such information is indispensable for syllabus design, it could also lead to a new kind of prescription, what we might call *frequency worship*, that is, concentrating on frequent items, patterns and structures at the expense of less frequent or idiosyncratic uses. Such practice deprives learners of alternative choices. Of course, it is not argued that learners should not be given frequency information, or that they should not be guided to become aware of the fact that some elements are more frequent than others, but that they should also be helped to realise that ‘less frequent’ does not mean ‘less acceptable’, and that ‘infrequent’ does not mean ‘wrong’. Similarly, learners should be made aware that frequencies change according to context of use.

The following analogy may help put the issue of frequency into perspective. Frequent items can be seen as the background, which is largely taken for granted, while the infrequent or idiosyncratic features foreground the user’s personality. Frequent items, used appropriately, help users blend in with a discourse community, whereas less frequent ones characterise individual language users. In view of this, language learners do need to be familiar with frequent features, which will enhance their understanding and production, but they should not be deprived of exposure to less frequent features, which will enable them to interpret nuances, enrich their own use, and help them express themselves in the new language. [-21-]

It is important to remember that concordance programs work with corpora, and that, consequently, the type and reliability of the derived information is contingent on the corpus that is used. Similarly, corpora are, ideally, representative samples of a language variety, a genre, or a medium (spoken or written). The misguided view of corpora as containing ‘the language’ may lead to generalisations from the examination of inappropriate corpora (e.g., generalising from a specialised corpus). Also, treating any large collection of texts as a corpus, that is, as a *representative* collection, may lead to conclusions based on the analysis of non-representative samples, for example, when using the Web as a corpus. This is not to deny that the Web is a vast, and freely available, resource of attested language use, but, rather, to stress that in order for the Web to be used effectively for teaching/learning purposes its users need to be aware of both its potential and limitations (see Kilgarriff & Grefenstette, 2003; Meyer et al., 2003; Robb, 2003; Volk, 2002). For example, the Web contains both NS and NNS English.

Finally, many language teachers have only limited access to corpora and corpus tools, usually through free online concordancers provided for demonstration purposes. These free tools allow for a small sample of concordance lines (usually 40-50), which may or may not be sufficient for learners to get a clear picture of the language feature they are investigating. These samplers usually give a fixed number of

words, typically 5-10, on either side of the key word/phrase, which may be inadequate for certain learning situations. Finally, these free tools do not always give information about the medium or genre of each concordance line. Also, because of limited data and restricted use of corpus software, teachers may see only easily observable patterns (e.g., adjacent collocations), and not less readily apparent ones (e.g., discontinuous collocations). Therefore, it would be wise to investigate any limitations of free corpus tools and take them into account.

Conclusion

Corpora and language description

Corpus-based linguistic research has provided increasingly clear and accurate descriptions of native and learner language, and has furnished linguistics and language teaching with new insights into language structure and use. Corpora have made it possible to compare native intuitions with actual use, and move from prescription to description. Thanks to corpora, language description for language teaching has been moving from over-generalised and exception-ridden rules towards flexible and context-specific patterns. Finally, due to corpus-based language analysis we are now in a position to identify the frequency of particular language features both with reference to language use as a whole and, more importantly, with reference to specific contexts. In fact, the analysis of large corpora not only makes it possible to identify frequent patterns and uses, but also affords enough data to examine rare or idiosyncratic ones. [-22-]

Corpora as language teaching tools

The increasing availability of corpora and ease of access to them, particularly through the World Wide Web, places a wealth of actual rather than made-up examples from different contexts at the fingertips of both teachers and learners. Corpus-based teaching is well suited to raising awareness of the varieties of English. Corpora also offer a welcome alternative to both specially-constructed pedagogical texts and authentic texts– the former being densely packed with the target language features, the latter offering only a partial picture of a language element. Another important contribution of corpora is the enhancement of discovery approaches to learning, which regard learners as language researchers. The development of corpus tools has also increased the value of the language lab.

Corpora, learners and teachers

The use of corpora in language teaching has helped redefine learner and teacher roles. It has reinforced learner-centred methodologies, and facilitated a further step away from the conception of teachers as sources of knowledge and providers of input, towards one of teachers as guides and facilitators, or even co-researchers. Corpus use has also introduced the need for learners and teachers to acquire new skills, and has placed increased emphasis on the necessity for teachers to develop their awareness of the language they teach. Finally, corpus-based research and teaching has the potential to empower non-native teachers and researchers, since native speaker introspection is no longer considered the one infallible source of insights into language structure and use.

Corpora and language teaching: What kind of relationship?

There is still a lot of ground to be covered until corpus use becomes a staple of language teaching and learning. In fact, if we wanted to describe the present relationship of most language teachers with corpora, then perhaps 'blind date' would be the most fitting metaphor. [22] However, the relationship between corpora and language teaching is definitely not 'a fling', as corpus-based materials and teaching approaches are becoming ever more pervasive in language teaching. But then again, it would be misleading to call the relationship a marriage, and short-sighted to wish it to become one. Corpora can and will continue to contribute greatly to language teaching in a multitude of ways, [23] but it would be misguided to treat them as a panacea. Corpus use is not meant to replace existing teaching methodologies, but to enrich and enhance them. If the time-dishonoured ELT pendulum is to be prevented from performing another one of its swings, then the use of corpora should not be treated as an alternative to, or rival of, existing teaching approaches, but as a welcome addition. [-23-]

References

- Alderson, J.C. (1996). Do corpora have a role in language assessment? In J.A. Thomas & M.H.Short (Eds.), *Using corpora for language research* (pp. 248-59). London: Longman.
- Altenberg, B. & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger (Ed.), *Learner English on computer* (pp. 80-93).

London: Longman.

Andrews, S. (1994). The grammatical knowledge/awareness of native-speaker EFL teachers: What the trainers say. In M. Bygate, A. Tonkyn & E. Williams (Eds.), *Grammar and the language teacher* (pp. 69-89). New York: Prentice Hall.

Arkin, E. 2003. *Teachers' attitudes towards computer technology use in vocabulary instruction*. Unpublished MA dissertation, Bilkent University.

Aston, G. (1996). The British National Corpus as a language learner resource. Paper presented at the second conference on Teaching and Language Corpora, Lancaster University, UK, 9-12 August 1996. Available online, <http://www.natcorp.ox.ac.uk/using/papers/aston96a.htm>.

Aston, G. (1997). Enriching the learning environment: Corpora in ELT. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and language corpora* (pp. 51-64). New York: Addison Wesley Longman.

Aston, G. (2000). Corpora and language teaching. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective: Papers from the third International Conference on Teaching and Language Corpora* (pp. 7-17). Hamburg: Peter Lang.

Aston G., Bernardini, S. & Stewart, D. (Eds.) (2004). *Corpora and language learners*. Amsterdam/Philadelphia: Benjamins.

Atwell, E., Gent, P., Medori, J. & Souter, C. (2003). Detecting student copying in a corpus of science laboratory reports. In D. Archer, P. Rayson, A. Wilson & T. McEnery (Eds.), *Proceedings of CL2003: International Conference on Corpus Linguistics*. UCREL technical paper number 16 (pp.48-53). UCREL, Lancaster University. Available online, <http://www.comp.leeds.ac.uk/eric/cl2003/AtwellGentMedoriSouter.doc>

Atwell, E., Howarth, P. & Souter, C. (2003). The ISLE corpus: Italian and German spoken learner's English. *ICAME Journal*, 27, 5-18.

Ball, F. (2001). Using corpora in language testing. *Research Notes* 6, 6-8, http://www.cambridgeesol.org/rs_notes/rs_nts6.pdf

Barker, F. (2004). Using corpora in language testing. *Modern English Teacher*, 13(2), 63-67.

Bekiou, K. & Díaz, L.(2004). What the use of Childes can say in analysing second language acquisition data: The acquisition of Spanish simple past tenses by Greek L1 learners. In Lewandowska-Tomaszczyk B. (Ed.), *Practical applications in language and computers (PALC 2003)* (pp. 321-342). Frankfurt: Peter Lang.

Bernardini, S. (2002). Exploring new directions for discovery learning. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis*. Proceedings from the Fourth International Conference on Teaching and Language Corpora, Graz 19-24 July, 2000 (pp. 165-182). Amsterdam: Rodopi. [-24-]

Bernardini, S. (2004). Corpora in the classroom: An overview and some reflections on future developments. In J.McH. Sinclair (Ed.), *How to use corpora in language teaching* (pp.15-36). Amsterdam: Benjamins.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8, 243-257.

Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, D. & Reppen, R. (1998). Comparing native and learner perspectives on English grammar: A study of complement clauses. In S. Granger (Ed.), *Learner English on computer* (pp. 145-158). London: Longman.

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

Bod, R. (1998). *Beyond grammar: An experience-based theory of language*. CSLI lecture notes 88. Stanford, CA: Centre for Study of

Language and Information Publications.

- Borg, S. (1999). The use of grammatical terminology in the second language classroom: A qualitative study of teachers' practices and cognitions. *Applied Linguistics*, 20(1), 95-126.
- Bradford, R. (2002). Grammar is by statisticians, language is by humans. *IATEFL Issues*, 167, 13. Available online, <http://www.tefl-china.net/ca23178.htm>.
- Carter, R. & McCarthy, M. (1996). Correspondence. *ELT Journal*, 50(4), 369-371.
- Close, R.A. (1992, rev. ed.) *A teachers' grammar: The central problems of English*. Hove, UK: Language Teaching Publications.
- Cobb, T. (1997). Is there any measurable learning from hands-on concordancing? *System*, 3(25), 301-315.
- Cohen, A. (2003, August). Strategy training for second language learners. *ERIC Digest*, <http://www.cal.org/resources/digest/0302cohen.html>.
- Coxhead, A. (2002). The academic word list: A corpus-based word list for academic purposes. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis* (pp. 73-89). Amsterdam: Rodopi.
- De Cock, S., Granger, S., Leech, G. & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learner English on computer* (pp. 67-79). London: Longman.
- Decoo, W. (2001). On the mortality of language learning methods. Article based on his L. Barker lecture at Brigham Young University, 8 November 2001. Available online, <http://www.didascalie.be/mortality.htm>.
- Dellar, H. (2003). What have corpora ever done for us? *Developing Teachers*, http://www.developingteachers.com/articles_tchtraining/corporapf_hugh.htm.
- Dudley-Evans, T. & St John, M.J. (1998). *Developments in English for Specific Purposes: A multi-disciplinary approach*. Cambridge: Cambridge University Press.
- Ferguson, G. (2001). If you pop over there: A corpus-based study of conditionals in medical discourse. *English for Specific Purposes*, 20, 61-82.
- Flowerdew, L. (2000). Investigating referential and pragmatic errors in a learner corpus. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective: Papers from the third International Conference on Teaching and Language Corpora* (pp. 145-154). Hamburg: Peter Lang. [-25-]
- Flowerdew, J. (Ed.) (2002). *Academic discourse*. Harlow: Longman.
- Fotos, S. & Ellis, R. (1991). Communicating about grammar: A task-based approach. *TESOL Quarterly*, 25(4), 605- 628.
- Fulcher, G. (1991). Conditionals revisited. *ELT Journal*, 45(2), 164-168.
- Gabrieli, R. (1998). Incorporating a student corpus in your teaching. *IATEFL Newsletter*, 141, 14-15.
- Gabrielatos, C. (1994a). Collocations: Pedagogical implications, and their treatment in pedagogical materials. Unpublished Master's essay, Research Centre for English and Applied Linguistics, Cambridge University, UK. Available online, <http://www.gabrielatos.com/Collocation.htm>.
- Gabrielatos, C. (1994b). Minding our Ps. *Current Issues*, 3, 5-8. Available online, <http://www.gabrielatos.com/MindingOurPs.htm>.
- Gabrielatos, C. (1999). Inference: Procedures and implications for TEFL. Part 1: Background. *TESOL Greece Newsletter*, 63, 15-20. A

unified and revised version was published in R.P. Millrood (Ed.) (2002). *Research methodology: Discourse in teaching a foreign language* (pp. 30-52). Tambov: Tambov State University Press, and is available through *ERIC* (ED476840), and online, <http://www.gabrielatos.com/Inference.htm>.

Gabrielatos, C. (2001). Shopping at the ELT supermarket: Principled decisions and practices. *ELT News*, 144. Revised version (2002) available through *ERIC* (ED478747), and online, <http://www.gabrielatos.com/ELTSupermarket.htm>.

Gabrielatos, C. (2002a). The Shape of the language teacher. In A. Pulverness (Ed.), *IATEFL 2002: York conference selections* (pp. 75-78). Whitstable, Kent: IATEFL. Available through *ERIC* (ED477571), and online, <http://www.ihes.com/ttsig/resources/articles/31.doc>.

Gabrielatos, C. (2002b). Grammar, grammars and intuitions in ELT: A second opinion. *IATEFL Issues*, 170, 2-3. Available online, <http://www.gabrielatos.com/Grammar-Intuitions.htm>.

Gabrielatos, C. (2003a). How much is that methodology in the window? *ESL MiniConference Online*, <http://www.eslminiconf.net/feb2003/gabrielatos.html>.

Gabrielatos, C. (2003b). Conditionals: ELT typology and corpus evidence. Paper given at 36th Annual Meeting of the British Association for Applied Linguistics (BAAL), University of Leeds, UK, 4-6 September 2003.

Gabrielatos, C. & McEnery, T. (2005, in press). Epistemic modality in MA dissertations. In P.A. Fuertes Olivera (Ed.), *Lengua y sociedad: Aportaciones recientes en lingüística cognitiva, lenguas en contacto, lenguajes de especialidad y lingüística del corpus* (pp. 311-331). Valladolid: Universidad de Valladolid.

Gavioli, L. (1997). Exploring texts through the concordancer: Guiding the learner. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and language corpora* (pp. 83-99). New York: Addison Wesley Longman.

Gledhill, C. (2000). The discourse function of collocation in research article introductions. *English for Specific Purposes* 19, 115-135.

Granger S. (Ed.) (1998). *Learner English on computer*. London: Longman. [-26-]

Granger, S. (1999). Use of tenses by advanced EFL learners: Evidence from an error-tagged computer corpus. In Hasselgard, H. & S. Oksefjell (Eds.), *Out of corpora. Studies in honour of Stig Johansson* (pp. 191-202). Amsterdam: Rodopi.

Granger, S. & Rayson, P. (1998). Automatic profiling of learner texts. In S. Granger (Ed.), *Learner English on computer* (pp. 119-131). London: Longman.

Granger, S. & Tribble, C. (1998). Learner corpus data in the foreign language classroom: Form-focused instruction and data-driven learning. In S. Granger (Ed.), *Learner English on computer* (pp. 199-209). London: Longman.

Granger, S., Hung, J. & Petch-Tyson, S. (Eds.) (2002). *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins.

Hahn, A. (2000). Grammar at its best: The development of a rule- and corpus-based grammar of English tenses. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective: Papers from the third International Conference on Teaching and Language Corpora* (pp. 193-206). Hamburg: Peter Lang.

Harwood, N. (2004). Citation analysis: A multidisciplinary perspective on academic literacy. In M. Baynham, A. Deignan, & G. White (Eds.), *Applied linguistics at the interface* (pp.79-89). London: Equinox.

Higgins, J. & Johns, T. (1984). *Computers in language learning*. London: Collins ELT.

Hughes, G. (1997). Developing a computing infrastructure for corpus-based teaching. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and language corpora* (pp. 292-307). New York: Addison Wesley Longman.

- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Hunston, S. & Francis, G. (1998). Verbs observed: A corpus-driven pedagogic grammar. *Applied Linguistics*, 19(1), 45-2.
- Hunston, S. & Francis, G. (1999). *Pattern grammar*. Amsterdam: John Benjamins.
- Hutchinson, T. & Waters, A. (1987). *English for specific purposes: A learning-centred approach*. Cambridge: Cambridge University Press.
- Hyland, K. (1999). Academic attribution: Citation and the construction of disciplinary knowledge. *Applied Linguistics* 20(3), 341-367.
- Hyland, K. (2002). Directives: Argument and engagement in academic writing. *Applied Linguistics* 23(2), 215-239.
- Hyland, K. & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing* 6(2), 185-205.
- James, K. & Garrett, P. (Eds.) (1991). *Language awareness in the classroom*. London: Longman.
- Johns, T. (1986). Microconcord: A language-learner's research tool. *System*, 14(2), 151-162.
- 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). Foris.
- Johns, T. (1991a). Should you be persuaded: Two examples of data driven learning. In T. Johns & P. King (Eds.), *Classroom concordancing*. ELR Journal 4 (pp. 1-16). Birmingham: University of Birmingham.
- Johns, T. (1991b). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Johns & P. King, P. (Eds.), *Classroom concordancing*. *ELR Journal*, 4 (pp. 27-45). Birmingham: University of Birmingham. [-27-]
- Johns, T. (1997). Contexts: The background, development and trialling of a concordance-based CALL program. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and language corpora* (pp. 100-115). New York: Addison Wesley Longman.
- Kennedy, G. (1992). Preferred ways of putting things with implications for language teaching. In Svartvik, J. (Ed.), *Directions in corpus linguistics*. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991 (pp. 335-378). Berlin: Mouton de Gruyter.
- Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Longman.
- Kennedy, C. and Miceli, T. (2001). An evaluation of intermediate students' approaches to corpus investigation. *Language Learning and Technology*, 5(3), 77-90. Available online, <http://llt.msu.edu/vol5num3/kennedy/default.html>.
- Kettemann, B. (1995). On the use of concordancing in ELT. *TELL & CALL*, 4, 4-15.
- Kilgarriff, A. & Grefenstette, G. (Eds.) (2003). *Web as corpus*. Special issue of *Computational Linguistics*, 29(3).
- Krashen, S.D. (1985). *The input hypothesis*. London: Longman.
- Krashen, S.D. (2004, 2nd ed.) *The power of reading: Insights from the research*. Portsmouth, NH: Heinemann.
- Leech, G.N. (1986). Automatic grammatical analysis and its educational applications. In G.N. Leech & C. Candlin (Eds.), *Computers in English language teaching and research* (pp. 205-214). Harlow: Longman.
- Leech, G. (1994). Students' grammar – teachers' grammar – learners' grammar. In M. Bygate, A. Tonkyn & E. Williams (Eds.), *Grammar and the language teacher* (pp. 17-30). New York: Prentice Hall.
- 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). New York: Addison Wesley Longman.

Lightbown, P. (1985). Can language acquisition be altered by instruction? In K. Hyltenstam & M. Pienemann (Eds.), *Modelling and assessing second language acquisition* (pp. 101-112). Clevedon, North Somerset: Multilingual Matters.

Lorenz, G. (1998). Overstatement in advanced learners' writing: Stylistic aspects of adjective intensification. In S. Granger (Ed.), *Learner English on computer* (pp. 53-66). London: Longman.

Loschky, L. & Bley-Vroman, R. (1993). Grammar and task-based methodology. In G. Crookes & S.M. Gass (Eds.), *Tasks and language learning: Integrating theory & practice* (pp. 123-166). Clevedon: Multilingual Matters.

Lucas, N., Cremilleux, B. & Turmel, L. (2003). Signalling well-written academic articles in an English corpus by text mining techniques. In D. Archer, P. Rayson, A. Wilson & T. McEnery (Eds.), *Proceedings of Corpus Linguistics 2003*. Technical Papers Vol. 16 – Special Issue (pp. 465-474). University Centre for Computer Corpus Research on Language (UCREL), Lancaster University.

Luzon Marco, M.J. (2000). Collocational frameworks in medical research papers: A genre-based study. *English for Specific Purposes* 19, 63-68.

Lyon, C., Barrett, R. & Malcolm, J. (2004). A theoretical basis to the automated detection of copying between texts, and its practical implementation in the Ferret plagiarism and collusion detector. Paper presented at Plagiarism: Prevention, Practice and Policies Conference, June 2004. Available online, <http://homepages.feis.herts.ac.uk/~comrcml/LyonPaperFerret.pdf>. [-28-]

Master, P. & Brinton, D. (Eds.) (1998). *New ways in English for specific purposes*. Alexandria, VA: Teachers of English to Speakers of Other Languages (TESOL).

Maule, D. (1988). "Sorry, but if he comes, I go": Teaching conditionals. *ELT Journal*, 42(2), 117-123.

McCarthy, M. (1990). *Vocabulary*. Oxford: Oxford University Press.

McEnery, T., Wilson, A. & Baker, J.P. (1997). Teaching grammar again after twenty years: Corpus-based help for teaching grammar. *ReCALL Journal*, 9(2), 8-17.

McEnery, T. & Wilson, A. (2001, 2nd ed.) *Corpus linguistics*. Edinburgh University Press.

McEnery, T. & Kifle N.A. (2002). Epistemic modality in argumentative essays of second-language writers. In J. Flowerdew (Ed.), *Academic discourse* (pp. 182-195). Harlow: Longman.

McEnery, T. & Gabrielatos, C. (2005, forthcoming). English corpus linguistics. In B. Aarts & A. McMahon (Eds.), *Handbook of English linguistics*. London: Blackwells.

McEnery, T., Xiao, Z. & Tono, Y. (2005, in press). *Corpus based language studies*. London: Routledge.

McKay, S. (1980). Developing vocabulary materials with a computer corpus. *RELC Journal*, 11(2), 77-87.

Meyer, C.F. (1991). A corpus-based study of apposition in English. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics. Studies in honour of Jan Svartvik* (pp. 166-181). London: Longman.

Meyer, C.F. (2002). *English corpus linguistics: An introduction*. Cambridge: Cambridge University Press.

Meyer, C.F., Grabowski, R., Han, H-Y., Mantzouranis, K. & Moses, S. (2003). The world wide web as linguistic corpus. In P. Leistyna & C.F. Meyer (Eds.), *Corpus analysis: Language structure and language use* (pp. 241-254). Amsterdam/New York: Rodopi.

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. 186-198). London: Longman.

- Mindt, D. (1997). Corpora and the teaching of English in Germany. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and language corpora* (pp. 40- 50). New York: Addison Wesley Longman.
- Nation, P. (1997). The language learning benefits of extensive reading. *The Language Teacher Online*, 21(5), 13-16, <http://www.jalt-publications.org/ilt/files/97/may/benefits.html>.
- Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In J.McH. Sinclair (Ed.), *How to use corpora in language teaching* (pp.125-152). Amsterdam: Benjamins.
- Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.
- Nivre, J., Allwood, J., Holm, J., Lopez-Kästen, D., Tullgren, K., Ahlsan, E., Gronqvist, L. & Sofkova, S. (1998). Towards multimodal spoken language corpora: TransTool and SyncTool. In *Proceedings of the Workshop on Partially Automated Techniques for Transcribing Naturally Occurring Speech at COLING-ACL '98*. Available online, <http://acl.eldoc.ub.rug.nl/mirror/W/W98/W98-0802.pdf>
- Nunan, D. (1989). *Designing tasks for the communicative classroom*. Cambridge: Cambridge University Press. [-29-]
- Nuttall, C. (1996, 2nd ed.). *Teaching reading skills in a foreign language*. Oxford: Heinemann.
- Osbourne, J. (2000). What can students learn from a corpus? Building bridges between data and explanation. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective*: Papers from the third International Conference on Teaching and Language Corpora (pp. 193-205). Hamburg: Peter Lang.
- Owen, C. (1993). Corpus-based grammar and the Heineken effect: Lexico-grammatical description for language learners. *Applied Linguistics*, 14(2), 167-187.
- Oxford, R. (1994, October). Language learning strategies: An update. *ERIC Digest*, <http://www.cal.org/resources/digest/oxfordo1.html>
- Partington, A. (1996). *Using corpora for English language research and teaching*. Amsterdam: John Benjamins.
- Pravec, N.A. (2002). Survey of learner corpora. *ICAME Journal*, 26, 81-114. Available online, <http://nora.hd.uib.no/icame/ij26/pravec.pdf>.
- Prodromou, L. (1996a). Correspondence. *ELT Journal*, 50(1), 88-89.
- Prodromou, L. (1996b). Correspondence. *ELT Journal*, 50(4), 371-373.
- Prodromou, L. (1997). Corpora: The real thing? *English Teaching Professional*, 5, 2-6.
- Rayson, P. (2001). Wmatrix: A web-based corpus processing environment. Software demonstration presented at ICAME 2001 conference, Universite Catholique de Louvain, Belgium. May 16-20, 2001. Available online, <http://www.comp.lancs.ac.uk/computing/users/paul/publications/icame01.pdf>.
- Rayson, P. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Unpublished PhD thesis, Lancaster University. Available online, <http://www.comp.lancs.ac.uk/computing/users/paul/phd/phd2003.pdf>.
- Read, C. (1985). Presentation, practice and production at a glance. In S.A. Matthews, M. Spratt & L. Dangerfield (Eds.), *At the chalkface: Practical techniques for language teaching* (p. 17). London: Edward Arnold.
- Renouf, A. (1997). Teaching corpus linguistics to teachers of English. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and language corpora* (pp. 255-266). New York: Addison Wesley Longman.
- Robb, T. (2003). Google as a quick 'n' dirty corpus tool. *TESL Electronic Journal*, 7(2), <http://www-writing.berkeley.edu/TESL-EJ/ej26/int.html>.

- Robinson, P.C. (1991). *ESP today: A practitioner's guide*. Hemel Hempstead: Prentice Hall.
- Rutherford, W.E. & Sharwood Smith, M. (1988). *Grammar and second language teaching*. New York: Newbury House.
- Schmidt, R.W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11(2), 129-158.
- Sharwood Smith, M. (1981). Consciousness-raising and the second language learner. *Applied Linguistics*, 2(2), 159-169.
- Sinclair, J.M. (1982). Linguistics and the teacher. In R. Carter (Ed.), *Linguistics and the teacher* (pp. 16-30). London: Routledge & Kegan Paul. [-30-]
- Sinclair, J.M. (1986). Basic computer processing of long texts. In G.N. Leech & C. Candlin (Eds.), *Computers in English language teaching and research* (pp. 185-203). Harlow: Longman.
- Sinclair, J.M. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Sinclair, J.M. (1997). Corpus evidence in language description. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and language corpora* (pp. 27-39). New York: Addison Wesley Longman.
- Sinclair, J.M. (2004). Intuition and annotation – the discussion continues. In K. Aijmer & B. Altenberg (Eds.), *Advances in corpus linguistics*. Papers from the 23rd International conference on English Language Research on Computerized Corpora (ICAME 23), Goteborg 22-26 May 2002 (pp. 39-59). Amsterdam: Rodopi.
- Skehan, P. (1993). A framework for the implementation of task-based learning. In *IATEFL 2003 annual conference report: Part 2, plenaries* (pp. 17-25). Whitstable, Kent: IATEFL.
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.
- Spratt, M. (1985). Presentation, practice, production. In S.A. Matthews, M. Spratt & L. Dangerfield (Eds.), *At the chalkface: Practical techniques for language teaching* (pp. 12- 16). London: Edward Arnold.
- Stern, H.H. (1992). *Issues and options in language teaching*. Oxford: Oxford University Press.
- Stevens, V. (1995) Concordancing with language learners: Why? When? What? *CAELL Journal*, 6 (2), 2-10. Available online, <http://www.ruf.rice.edu/~barlow/stevens.html>.
- Stubbs, M. (1996). *Text and corpus analysis: Computer-assisted studies of language and culture*. Oxford: Blackwell.
- Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.
- Susser, B. & Robb, T.N. (1990). EFL extensive reading instruction: Research and procedure. *JALT Journal*, 12(2), <http://www.kyoto-su.ac.jp/~trobb/sussrobb.html>.
- Swales, J.M. (Ed.) (1985). *Episodes in ESP*. Oxford: Pergamon.
- Swan, M. (1994). Design criteria for pedagogic language rules. In M. Bygate, A. Tonkyn & E. Williams (Eds.), *Grammar and the language teacher* (pp. 45-55). New York: Prentice Hall.
- Taylor, D. (1994). Inauthentic authenticity or authentic inauthenticity? *TESL Electronic Journal*, 1(2), A-1, <http://writing.berkeley.edu/TESL-EJ/ejo2/a.1.html>.
- Thompson, P. (2002). Modal verbs in academic writing. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis* (pp. 305-325). Amsterdam: Rodopi.

- Thompson, P. & Tribble, C. (2001). Looking at citations: Using corpora in English for academic purposes. *Language Learning & Technology* 5(3), 91-105.
- Thornbury, S. (2000). A dogma for EFL. *IATEFL Issues*, 153, 2. Available online, <http://www.teaching-unplugged.com/dogmaarticle.html>.
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins. [-31-]
- Tono, Y. (2000). A computer learner corpus based analysis of the acquisition order of English grammatical morphemes. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective*. Papers from the third International Conference on Teaching and Language Corpora (pp. 123-132). Hamburg: Peter Lang.
- Tribble, C. (1997a). Improvising corpora for ELT: Quick-and-dirty ways of developing corpora for language teaching. In J. Melia. & B. Lewandowska-Tomaszczyk (Eds.), *PALC '97 Proceedings, Practical Applications in Language Corpora* (pp. 106-117). Lodz: Lodz University Press. Available online, <http://www.ctribble.co.uk/text/Palc.htm>.
- Tribble, C. (1997b). Put a corpus in your classroom: Using a computer in vocabulary development. In T. Boswood (Ed.), *New ways of using computers in language teaching* (pp. 266-268). Alexandria, VA: TESOL.
- Tribble, C. (2000). Practical uses for language corpora in ELT. In P. Brett & G. Motteram (Eds.), *A Special interest in computers* (pp. 31-42). Whitstable, Kent: IATEFL.
- Tribble, C. & Jones, G. (1990) *Concordances in the classroom*. London: Longman.
- van Halteren, H. (2003). Detection of plagiarism in student essays. Paper presented at the 14th meeting of Computational Linguistics in the Netherlands, University of Antwerp, Belgium, 19 December 2003. Available online, <http://cnts.uia.ac.be/clin2003/proc/13Vanhalter.pdf>.
- Virtanen, T. (1997). The Progressive in NS and NNS student compositions: Evidence from the International Corpus of Learner English. In M. Ljung (Ed.), *Corpus-based studies in English*. Papers from the Seventeenth International conference on English Language Research on computerized Corpora (ICAME 17) (pp. 299-309). Amsterdam: Rodopi.
- Virtanen, T. (1998). Direct questions in argumentative student writing. In S. Granger (Ed.), *Learner English on computer* (pp. 94-106). London: Longman.
- Volk, M. (2002). Using the web as corpus in linguistic research. In R. Pajusalu & T. Hennoste (Eds.), *Tahendusepuudja. Catcher of the meaning. A festschrift for Professor Haldur Oim*. Publications of the Department of General Linguistics 3. University of Tartu. Available online, http://www.ifi.unizh.ch/cl/volk/papers/Oim_Festschrift_2002.pdf.
- Wallace, C. (1992). *Reading*. Oxford: Oxford University Press.
- Wang, S. (1991). *A corpus study of English conditionals*. Unpublished MA thesis, Victoria University of Wellington.
- Westney, P. (1994). Rules and pedagogical grammar. In T. Odlin (Ed.), *Perspectives on pedagogical grammar* (pp. 72-96). Cambridge: Cambridge University Press.
- Wible, D., Kuo, C-H., Chien, F. & Wang, C.C. (2002). Toward automating a personalized concordancer for data-driven learning: A lexical difficulty filter for language learners. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis*. Proceedings from the Fourth International Conference on Teaching and Language Corpora, Graz 19-24 July, 2000 (pp. 147-154). Amsterdam: Rodopi.
- Widdowson, H.G. (1979). *Explorations in applied linguistics*. Oxford: Oxford University Press.
- Widdowson, H.G. (1990). *Aspects of language teaching*. Oxford: Oxford University Press.
- Widdowson, H.G. (1991). The description and prescription of language. In J. Alatis (Ed.), *Georgetown University Round Table on Languages and Linguistics 1991* (pp. 11-24). Washington, D.C.: Georgetown University Press. [-32-]

Willis, J. (1996). *A framework for task-based learning*. Oxford: Longman.

Xiao, Z. (forthcoming). Well-known and influential corpora. In A. Ludeling, M. Kyto & A.M. McEnery (Eds.), *Corpus linguistics: An international handbook. Handbooks of linguistics and communication science*. Berlin: Mouton de Gruyter.

[-33-]

Notes

[*] This paper is based on my plenary address at INGED 2003 International Conference, *Multiculturalism in ELT Practices: Unity and Diversity*, organised jointly by BETA (Romania), ETAI (Israel), INGED/ELEA (Turkey) and TESOL Greece, Baskent University, Ankara, Turkey, 10-12 October 2003. I would like to thank Paul Baker (Lancaster University) for directing me to relevant corpus studies. Special thanks are due to Alikı Chapple for editing the different incarnations of this paper.

[1] For short overviews see McEnery & Gabrielatos (2005, forthcoming), Pravec (2002).

[2] See also the debate in the *Correspondence* section of *ELT Journal* (Carter & McCarthy, 1996; Prodromou, 1996a, 1996b)

[3] For a comprehensive account of native-speaker and learner corpora, with relevant references and links to websites, see Xiao (forthcoming), <http://www.lancs.ac.uk/postgrad/xiaoz/papers/corpus%20survey.htm>. For a glossary of terms used in corpus linguistics see McEnery & Wilson (2001).

[4] Of course, corpus compilers need to have already secured permission from the copyright holders.

[5] For a discussion of mark-up and annotation see McEnery et al. (2005).

[6] The annotation was carried out automatically by the *Wmatrix* interface (Rayson, 2001, 2003), <http://www.comp.lancs.ac.uk/ucrel/wmatrix/>, using the *CLAWS* part of speech tagger, <http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/>. For a guide to the complete tagset used in the BNC see <http://www.comp.lancs.ac.uk/ucrel/bnc2/bnc2guide.htm>

[7] The initial sample was 1,000 sentences, but was reduced to 831 after non-conditional uses of *if* and sentences with *even if* were removed. As the coursebooks examined presented only syntactically straightforward sentences, and in order to examine ELT materials on their own terms, cases of embedded or elliptical clauses and idiomatic uses were also excluded, leaving 710 *if*-conditionals.

[8] See also Ferguson (2001), Fulcher (1991), Maule (1988), Wang (1991).

[9] Although EAP is a sub-field of ESP, as both terms refer to language teaching with a better defined and narrower focus than “general English,” it has become common practice to treat ESP and EAP as somehow distinct, though closely related, areas (see Dudley-Evans & St. John, 1998; Hutchinson & Waters, 1987; Masters & Brinton, 1998; Robinson, 1991; Swales, 1985).

[10] See also the report of the Learning Technologies Group, Oxford University Computing Services (http://www.oucs.ox.ac.uk/ltg/reports/plag_index.xml).

[11] On the use of terminology in language teaching see Borg (1999). [-34-]

[12] For example, identifying the topic, the gist, or specific information, or using the context, co-text and background knowledge to infer meaning or attitude (see Nuttall, 1996; Wallace, 1992).

[13] Calculated on the basis of a standard A4 page, single spaced, using Times New Roman 12.

[14] The term “authentic texts” is used here as a shorthand for “texts addressed to native speakers of a language.” For a discussion of “authenticity” in language teaching see Taylor (1994), Widdowson (1979, 1990).

[15] *Text-based* refers to the use of a single text, or a small number of short texts, as language data. For examples of text-based language awareness studies see James & Garrett (1991).

[16] The texts used were: “Diet Guidelines Aimed at Healthy People,” by Emily Gersema, 26 September 2003, http://www.stopgettingsick.com/templates/news_template.cfm/7061 “Quick-fix diets fail fat Britons,” by Jo Revill, *The Observer*, 5 January 2003, http://observer.guardian.co.uk/uk_news/story/0,6903,868891,00.html; “You CAN Lose the Weight You Want!” <http://www.eatandburn.com/3/?rid=22&code=ncdfr20607&publisher=&transid=>.

[17] The concordance was derived from the three texts using *Wmatrix* (see note 6).

[18] Sample 4 is given in an alternative view to a concordance, called “sentence view.” Both sets of data are from the British National Corpus and were derived using the *BNCweb* interface, developed at the University of Zurich (<http://homepage.mac.com/bncweb/home.html>).

[19] From this point on, “corpus sample” will be used to refer to collections of examples from a corpus in either “concordance” or “sentence” format.

[20] For reasons of space, only a random sample of 10 sentences is given here; students should be given a larger corpus sample so that more patterns are discernible.

[21] For a discussion of different views on the role of theory and intuitions in corpus-based research see McEnery & Gabrielatos (2005, forthcoming).

[22] I would like to thank Alan Waters (Lancaster University) for prompting this metaphor by suggesting “first date.”

[23] For example, the development and availability of *multimodal spoken corpora*, that is, corpora in which the transcribed text is linked to sound and video files (Nivre, et al., 1998), will enable researchers, materials writers, teachers and learners to use corpora to focus on phonological features, as well as facial expressions, gestures and body language.

About the Author

Costas Gabrielatos has worked as a language teacher, teacher educator and lecturer. He is currently a doctoral candidate at Lancaster University, UK, doing corpus-based research on English if-conditionals. He is also a part-time tutor and researcher in linguistics. His main areas of interest are tense, aspect and modality in English, and the implications of corpus-based language analysis for language teaching and teacher education.

Appendices

Appendix 1. Free/affordable corpora and corpus tools

- British National Corpus Sampler (1 million words of written and 1 million words of spoken English): <http://www.natcorp.ox.ac.uk/getting/sampler.html>. Also, free, but restricted, access to the full BNC: <http://sara.natcorp.ox.ac.uk/lookup.html>
- Collins Wordbanks *Online* English corpus (concordance and collocation samplers): <http://www.collins.co.uk/Corpus/CorpusSearch.aspx>
- The Compleat Lexical Tutor: <http://132.208.224.131>
- Michigan Corpus of Academic Spoken English (MICASE): <http://www.hti.umich.edu/m/micase>
- Variation in English Words and Phrases (Mark Davies, Brigham Young University). Interface to the full British National Corpus (100 million words): <http://view.byu.edu/>
- Web Concordancer (works with a variety of corpora): <http://www.edict.com.hk/concordance/>
- WebCorp: The Web As Corpus (University of Liverpool): <http://www.webcorp.org.uk/>
- WordNet: A Lexical Database for the English Language (Princeton University): <http://www.cogsci.princeton.edu/~wn>
- WordSmith Tools: <http://www1.oup.co.uk/elt/catalogue/Multimedia/WordSmithTools3.0>

Appendix 2. Online courses and information on language corpora

- Concordancing in the classroom / Data-driven learning techniques: <http://www.edict.com.hk/concordance/aboutweb.htm#Concordancing%20in%20the%20classroom>
- Corpus Linguistics: A Practical Web-based Course (Lancaster University): http://www.ling.lancs.ac.uk/courses/ahaw-nsl/clc_top.htm
- A companion to *Corpus Linguistics* (McEnery & Wilson, 2001): <http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/contents.htm>
- Information and Communications Technology for Language Teachers, Module 3.4, Corpus Linguistics (written by Tony McEnery and Andrew Wilson): http://www.ict4lt.org/en/en_mod3-4.htm.
- Information and Communications Technology for Language Teachers, Module 2.4, Using concordance programs in the modern foreign languages classroom (written by Marie-Noëlle Lamy, Hans Jørgen Klarskov Mortensensium and Graham Davies): http://www.ict4lt.org/en/en_mod2-4.htm#_Toc481294151.
- Catherine Ball, Georgetown University: <http://www.georgetown.edu/faculty/ballc/corpora/tutorial.html>

© Copyright rests with authors. Please cite TESL-EJ appropriately.

Editor's Note: Dashed numbers in square brackets indicate the end of each page for purposes of citation.

