

7 *Corpora and language teaching: General applications*

In this chapter, applications of corpora to the practice of language teaching (mainly English language teaching) will be considered. There are sections devoted to data-driven learning and reciprocal learning, and to issues relating to methodology and syllabus design. The final section is a discussion of recent challenges to the use of corpora in language teaching.

Data-driven learning

Introduction

As Leech (1997c: 3) comments, the use of corpora in language teaching situations owes much to the work of Tim Johns, who developed data-driven learning (DDL) for use with international students at the University of Birmingham. An often-quoted comment by Johns is that ‘Research is too important to be left to the researchers’ (1991: 2). The theory behind DDL is that students act as ‘language detectives’ (Johns 1997a: 101), discovering facts about the language they are learning for themselves, from authentic examples. This supports learning, partly because students are motivated to remember what they have worked to find out. In addition, because corpus data can reveal previously unnoticed patterns, a student may well notice something that a teacher has overlooked, or that no textbook covers. As well as being beneficial in teaching specific items, DDL is hypothesised to improve general skills of using context to deduce meaning.¹ DDL involves setting up situations in which students can answer questions about language themselves by studying corpus data in the form of concordance lines or sentences. The questions may arise out of something the student is writing, and may be formulated as ‘Is it better to say x or y?’ or ‘What is the difference between saying x and saying y?’ In this case, the questions are the student’s own. Alternatively, self-access materials may be written that

¹ The hypotheses about the benefits of DDL have not yet been adequately tested, but see Stevens (1991), Cobb (1997) and Cobb and Horst (2001) for small-scale studies.

allow students to explore general issues such as ‘that-clauses’, using information from a corpus. In this case, the teacher aims to teach items which are known to be problematic or useful for the groups of students concerned. The first kind of study will use a ‘raw corpus’, in the sense that the student and tutor will look at the corpus together, without either of them necessarily knowing what they will find. For the second kind of study, the tutor has to carefully select and possibly edit the concordance lines in order to demonstrate the target language feature.

The first kind of study has the advantage of maximum student motivation: the student asks a question for which an answer is urgently required (for the student to complete a piece of written work, for example), and is therefore highly motivated to discover information in the corpus data consulted. A possible disadvantage for the teacher is that they have very little control over what happens. If the corpus is consulted and no answer is apparent to student or teacher, or if further difficult questions are raised, the teacher may feel that a loss of expertise has occurred. A more basic problem is that not every teaching situation allows the luxury of one-to-one consultations, or sufficient computer access for students to undertake investigations on their own. In the second kind of study, the teacher, having selected the information, has more control. Materials can be printed on to paper to be used with a whole class. The disadvantage is that, as the teacher has selected the topic for study, the students are potentially less motivated to search for or remember the target information. In these circumstances, DDL may appear to the students to be a tangential activity to the main business of the class.

More recent developments in data-driven learning (e.g. Bernadini 2000) stress the benefits of encouraging students to design their own corpus investigations and to take advantage of the ‘serendipity’ effect of searching a corpus when the agenda is not too firmly fixed and a student can follow up any interesting observations that they happen across. This Discovery Learning, as it is sometimes called, is most suitable for very advanced learners who are filling in gaps in their knowledge rather than laying down the foundations. At the other end of the scale, Cobb and Horst (2001) describe an experiment to teach large amounts of vocabulary to EAP students, using concordances from a corpus of texts from the students’ language course. In this very controlled environment, students learned lists of vocabulary items more successfully when they had access to the concordance lines than using other methods.

DDL with a 'raw' corpus

Advanced learners can safely be encouraged to use a raw (unedited) corpus to make observations about the language. Dodd (1997), for example, describes advanced learners of German using a corpus of German newspapers. Among other activities, the students test out statements made in standard reference books, about grammar, such as rules for the use of particular conjunctions, and lexis, such as differentiation between near-synonyms. The newspaper corpus can also be used to test hypotheses about the use of various key terms in East and West Germany. Many teachers nowadays use the worldwide web to allow students access to a range of corpora, both monolingual and parallel (Foucou and Kübler 2000) or to encourage students to build their own corpora (Pearson 2000).

The challenge for the teacher who wishes to encourage students to do work of this kind is to formulate a task in such a way that the student will obtain maximum benefit from it. If teacher and student are in a one-to-one consultation, the teacher can 'play it by ear'. In other circumstances, however, the teacher will have to do some planning of the activity to be undertaken. Here are two examples of activities which a student may undertake, using an unedited version of the Bank of English corpus.

Example 1: prevent

A student writes the following phrase in an essay: . . . *in their efforts to prevent such incidents to ever happen again*. The teacher disagrees with this use of *prevent*, and sends the student to a corpus to investigate the conventional usage of this verb. There are two alternative ways of doing this. One is to focus on the verb *PREVENT* and its patterns. The student can be asked to obtain 30 random lines from the corpus for the verb. This will yield useful information, but it does not reflect how the student probably built up the problem sentence, in which the focus could have been on *incidents* and the notion of 'happening again' as much as on *prevent*. If the corpus used is large enough, the student can be instructed to search for *PREVENT* and *INCIDENT* together. Here are 23 concordance lines from the Bank of English for each of these searches:

PREVENT

of the whole human family, has to prevent the horrific toll of 40,000
48, who was stabbed as he tried to prevent a gang attacking William Njoh, 13.
service authorities did nothing to prevent the dangerous material entering

for those whose work commitments prevent them from following the normal they must be turned regularly to prevent the embryo sticking to the shell simple erm and are designed to prevent radon getting from the ground up the only solution. And attempts to prevent them from leaving may create an out that only one man could now prevent war in the Gulf – and that man is may consider tabling legislation to prevent unmerited pay rises. The process. No one who is concerned to prevent the disintegration of an sick. The state had no authority to prevent the plant from burning chlorinated States Supreme Court in an effort to prevent Klansmen from striking all the seek help, and seek it quickly to prevent further damage to them. schism to extend our power. And to prevent the Cybernetic Universal Church social forces will by themselves prevent excessive mongrelization from to attract a mate, failing to prevent cuckoldry, or failing to keep a was content to use its influence to prevent any Senate amendments adding objectives in Louisiana: first, to prevent New Orleans from turning into an that he has legal rights which prevent such demeaning, mix and match bundle the leaves in the winter to prevent snow and water getting into the own way. A lone Labor voter prevented a clean sweep in Birdsville, V.P. Singh if the temple project is prevented or if the pilgrimage by its at least some batterings could be prevented, but now local advocacy groups

PREVENT . . . INCIDENT

have prevented. Talking about the incident became, in its own way, a part of predict or prevent that damage. The incident spoiled my self-satisfaction, oxygen system to **prevent a similar** incident. Verdict: Natural causes. should help **prevent further** incidents of the kind that Mr Rowe so The MPs hope to **prevent** incidents such as the attack on Daniel in an attempt to **prevent** potential incidents of ‘bar rage’. Plain- clothes would help **prevent a repeat of** incidents like the collapse of the State to act to **prevent such** violent incidents **again**.” Local MLA Vince and thought to **prevent such** incidents **in future**, Itar-Tass news agency should take steps ‘to **prevent such** incidents as the assassination of Rabin’. to have been unable to **prevent such** incidents. The fundamentalists who patrol two countries to **prevent similar** incidents **in the future**. DOMINICAN concerned to **prevent further such** incidents. The Secretary General of the decisive measures to **prevent** these incidents. Simon Long reports from Peking: could do far more to **prevent such** incidents. However the fact that this responsibility to **prevent such** incidents. Yesterday, the Lebanese to be failsafe and **prevent such** incidents from happening. It is a spy Rudolf Abel. To **prevent** incidents **like this**, air forces in the US guidelines which can help **prevent** incidents **like this one**. In the Persian Post Office can do to **prevent such** incidents. FRED VAN DE PUTTE, Postal was in its own interest to **prevent** incidents from occurring on the premises, To **prevent** terrorist incidents and to convict those responsible will be enough to **prevent any more** incidents. I don’t think Eric ever

What the teacher ‘wants’ the learner to see in these lines are the presence of the patterns *prevent something*, *prevent something happening* and *prevent something from happening* and the absence of the pattern *prevent something to happen*. The teacher may also wish the learner to note ways of expressing recurrence, as in *prevent*

incidents like this, prevent such incidents or prevent a similar incident. In other words, there are two foci of attention: the patterns of the verb and the ways of expressing the whole idea.

Once the concordance lines have been obtained, the learner may be asked simply to notice the patterns. With very advanced students, or students who have worked with concordance lines before, this may be a successful strategy. With other students, more detailed instructions may have to be given. For example, the student may be given these pattern phrases – *prevent an incident, prevent an incident happening, prevent an incident from happening* and *prevent an incident to happen* – and be asked to match each pattern phrase with concordance lines. This would draw the student's attention to what is relevant to pattern in the sample lines. If lines with *incident* have been selected, the student can also be instructed to underline, for example, all words and phrases indicating 'happen again'. Finally, the student can be asked to re-word the problem sentence in one or more ways.

Example 2: not any and no

This task replicates Dodd's idea of asking learners to check grammar book information in a corpus. The grammar book used is the *Longman English Grammar* (1988: 93–94) which notes that negatives can be formed using *not . . . any* or with *no*, as in *There aren't any buses after midnight* and *There are no buses after midnight*. A useful learner project might be to check a corpus for any differences between these alternative phraseologies. The corpus used here is the spoken corpus from the Bank of English. The first thing to be noticed is the difference in frequency. In this corpus, *there's no* is much more frequent than *there isn't any* (over 2,000 lines compared with fewer than 100). (This finding concurs with Biber et al 1999: 172.) Secondly, *there isn't any* frequently occurs at the end of a clause, with *any* being used cohesively, referring back to something earlier in the discourse, as in these (slightly edited) examples:

A: I mean what would you want them to do to help themselves?

B: Well perhaps look round for accommodation instead of just sitting on the street.

A: But there isn't any.

I dunno whether it's maybe just that the subtext of the Archers is much more clumsy or that there isn't any.

Take my point about there is no trust for the police. There isn't any.

There's nowhere within a few minutes walk of here that's selling sweets. And I've just been on the prowl to find somewhere and there isn't any.

A: What's the point of that?

B: There isn't any.

He said Can't we have some sprouts. I said No there isn't any.

A: What about the state of the streets litter er garbage collection that kind of thing?

B: There isn't any.

Apart from that, the phrase is used with a variety of nouns, as in these lines:

we go into school <M01> There isn't any **bus** that comes that way is category but there there isn't any clear **perception** that say in here is it? No. <M01> Cos there isn't any **decrease** in the number of kids the seventeenth of February. Now there isn't any **fiddle** going on because the it was probably because there isn't any free **parking** is there there War of course illustrates that there isn't any **government** at that level it's kind of because there isn't any **hot water** nothing got often things like noise and dust there isn't any **limit** or isn't any compliance different ways. <M0X> Mm I bet there isn't any **newspaper** in another language use that argument both ways. There isn't any one **reason** why it's become there isn't anything. There isn't any **other word** so if you had to <F01> Yeah. <M0X> Erm. <F01> There isn't any **possibility** I mean this is on Villa to sell although there isn't any **pressure** if erm if Doug takes <F01> And it's so unfair and there isn't any **redress** <F06> No that's you're doing a spoken one I mean there isn't any **sensible way** of sampling So I think you I think <F06> But there isn't any **such thing** yet is there John? is pressing her <M01> Mhm. <F01> there isn't any **time** you see this is the train of thought really because there isn't any **train of thought** very much. re actually bound booklets so there isn't any **V A T**. I don't think so re saying FX? Yes because there isn't any **water** under the rim.

The phrase *there's no*, on the other hand, is frequently followed by abstract, discourse nouns such as *need*, *point*, *problem*, *reason* and *way*, rather than nouns referring to physical objects. The following lines illustrate this:

know between as far as I know there's no **comparison** between the nature to make the point that there's no **evidence** whatsoever that humans can also be some colour although there's no **example** of it in this er particular actually. <M01> Mm. Well there's no **hurry** is there. <M02> No. is no <M01> Mm. <M02> er there's no **measure** of cloud liquid water in the it down in writing so that there's no er **misunderstanding** <M01> Mm half a dozen times. You know there's no **need** for it because by the time it at the end of the day there's no **need** for any er people say you know I'm not going at seven. There's no **point** in having a wide receiver so was there's there's no **point** in actually looking and sort arrest <M01> Mm. <F01> erm there's no **problem** with it but er I mean provi so you're already identified there's no **problem** at all really. <M01> No of s got a key and he gets in and there's no **problem**. So I think it might be my that's how it works and there's no **reason** why it shouldn't work like Mm. <F0X> starter. <F0X> There's no **sign** of them. <F0X> That's mine in from the courts they said so there's no **use** even putting in for it 'cos you

I miss a lot of things and there's no way the patient can ring me as well
 I hadn't got access to a car there's no way I could cart up bottles and <M01>
 s <M01> Yeah. <M02> there's no way of knowing precisely which detail
 I'll have to <FOX> Yeah because there's no way in which – You've got to teach

This small investigation illustrates the general rule that details of usage can account for observations of frequency and can make those observations more directly useful. This in turn illustrates the need to move between quantitative and qualitative information.

As was noted in chapter 1, Owen (1996) warns that encouraging advanced students to look in a corpus can lead to problems. He notes that a native speaker would probably find a sentence such as *Further experiments require to be done* unacceptable and might direct the student to a corpus in order to discover that *REQUIRE* is not used with a passive infinitive, whereas *NEED* is. The problem is that the student looking at the relevant concordance lines in the Bank of English would actually find a number of lines that appear to break the rule, such as these:

that at some stage you would require to be admitted to hospital for
 must raise other questions which require to be answered. Among them, he
 and women in these hospitals did not require to be cared for in such secure
 of the modifications that we require to be carried out on the yeast
 of the minority of patients who require to be detained in hospital or
 at the commercial potential that may require to be developed at a number of
 <p> Where media and cultural studies require to be distinguished is in their
 me as Official Receiver, would require to be funded. I have approached
 drugs are available. However, they require to be given for rather a prolonged
 been derived are too well known to require to be indicated. Countless letters
 require maintenance but it does require to be kept well moist and
 you can take unstable vaccines that require to be kept in the refrigerator and
 model. These may be prisoners who require to be kept apart for their own
 small air inclusions which will require to be made good in this way during
 legal position of the monarchy would require to be made by legislation enacted
 certain needs and goals that they require to be met <M01> Right
 that a large number of laws would require to be passed by a two-thirds
 <p> A Yes, your cordon pears do require to be pruned in summer. This
 as the physical health of the people require to be remedied.
 the fact that these prisoners require to be segregated for their own

In other words, the corpus information does not seem to accord with native-speaker intuition. Looking at the corpus may therefore confuse the student and undermine the authority of the teacher. Closer examination of lines such as those above suggests a resolution of the problem, in that in each line the past participle is of a verb indicating a specific action (e.g. *vaccines require to be kept*, *pears require to be pruned*, *prisoners require to be segregated*) rather than of a general verb such as *do*. In each case the subject of the clause

indicates an entity which is the goal of a process. The problem sentence *Further experiments require to be done* is different, not only because the verb is a general one but also because the subject expresses the range of the verb (in Halliday's terms) rather than the goal. The phrase *do an experiment* expresses a single action rather than an action done to something. Although this may solve Owen's dilemma, it is true that his question 'How many instances in a corpus are enough to show that something is correct English?' becomes a pressing, and awkward, one when learners investigate a corpus for themselves. An extreme example of this is clauses introduced by *like*, such as *They head for me like I'm a magnet or something*. This is usually considered to be 'incorrect' in English, the correct version being *They head for me as if I was a magnet or something*. The Bank of English has numerous examples of the incorrect usage, which is very common, especially in spoken English. Distinguishing between what is said and what is accepted as standard may need the assistance of a teacher or a grammar book.

Designing materials based on corpus data

The alternative to encouraging learners to explore a raw corpus is to select the evidence, that is, to give learners materials based on concordance lines which the teacher has selected, and to add questions which will guide the learners towards noticing relevant information in the lines. Johns (1997a: 101) gives some examples of question types (e.g. 'How many different verbs are shown with this structure?' and 'Which word is present in the right context of citations 1–8 that is not present in the right context of citations 9–16?') and notes that '[s]uch tasks are, of course, "closed" in the sense that the result is known to the teacher in advance'. The advantage of selecting concordance lines is that lines with exceptionally difficult vocabulary can be left out, as can lines that exemplify usages that the teacher would prefer the student to ignore at this stage. This selection allows concordance lines to be used with students who are not advanced enough to benefit from 'raw' concordance data.

The teacher may choose to begin with a word that is already familiar to learners. The following lines illustrate one use of the adjective *angry*:

At first I thought her parents were angry with her
 But you get so angry with me!
 how can you be angry with the man you love
 Ian gets angry with the television sometimes

These lines have been selected because they do not contain difficult language and because there is a whole sentence or clause in one line. The lines have been cut so that there is no extraneous information in them. Learners, having identified the sequence *angry with* as the key point here, can be asked to look at other lines with a similar pattern, such as:

I'm more annoyed with myself you know
 I was annoyed with him.
 I knew my father would be annoyed with me.
 But Americans also are annoyed with George Bush
 He says he's never been bored with the job
 I got a bit bored with popular music
 I get bored with cooking
 She's highly intelligent and gets bored with television
 I was clearly becoming rather impatient with rejections.
 Charlie could be impatient with others.
 There are countless times when I get impatient with my husband, Ken
 increasingly impatient with the slow pace of change

A simple exercise is to ask learners to list the adjectives used in this way. If they are able, they might be asked to predict other adjectives that they might expect to have the same pattern, and these can be checked against a corpus, or against a dictionary. Further exercises could include underlining the verbs that come before the adjectives (*BE, GET, BECOME*). If the concordance lines are extended, learners can note a longer phraseology, such as prepositional phrases beginning with *for* that express the reason for anger, as in these extended lines:

Was he not even angry with his mother for not explaining things t
 I do not feel angry with him for what he has done.
 I feel angry with her for not standing up to him.
 looked at him as if she were annoyed with him for letting it happen.

Learners are building up an extended phrase here that might be expressed as 'be angry with someone for something' or 'be angry with someone for (not) doing something'.

Exercises such as this one can be enjoyable, but they are also time-consuming for the teacher to write, so it is worthwhile bearing in mind this caveat from Dave Willis (personal communication):

One of the major problems with DDL, or with consciousness-raising in general, is what to focus on. Exercises of this kind are very time consuming in the classroom. A sequence like that starting with *angry* and leading on to *angry with someone for something* might not repay the time taken. . . . I would not argue that exercises like this are not worth doing. I am saying that time in classrooms is very limited and that there is a danger of spending too long on generalisations which may be of limited value . . .

To avoid a wasteful expenditure of effort, writers of DDL materials frequently focus on items which are known to be difficult for students with a particular language background, or items which are particularly frequent or otherwise important in a given subject area.

DDL exercises can be integrated into the rest of the lesson if the starting point for the activity is a word or phrase met in a reading or listening text, or in another classroom activity. For example, a reading passage with a group of students studying science through the medium of English may contain the sentence *Salt water has a lower freezing point than normal water*. The teacher may consider the phraseology of the phrase *freezing point*, and its close relatives *boiling point*, *melting point* and so on, to be important enough to warrant the development of DDL materials. It is a simple matter to select some concordance lines:

soil to raise the temperature above freezing point, the planetary permafrost
 hunt. Temperatures remained at freezing point, prompting Kobe doctor
 As the temperature dropped below freezing point at night, local reside
 drops in your area below freezing point for seven consecutive
 water, still liquid below its freezing point, and in a false state
 freezer since alcohol has a lower freezing-point than water and therefore
 imparting taste. They lower the freezing point of a food to keep it l
 on runways and aircraft, push the freezing point of ice down to 13 degrees
 increased pressure lowers the freezing point of water. This is the
 cold – don't you know what the freezing point of alcohol is? As we s
 the period never rose above the freezing point of water (32F). Certainly
 to car antifreeze, to lower their freezing point, to prevent large ice
 and play in temperatures near to freezing point. That's not a whinge
 that, sometimes nearly down to freezing point on a cold night. And s
 reduce the temperature inside to freezing point in seconds. 'Imagine

Questions for students could include the following:

- Underline the lines where *freezing point* does not have *a* or *the* or *its* or *their* in front of it. In the lines, which words come before freezing point?
- Complete these sentences:
The temperature dropped freezing point.
The temperature rose freezing point.
The temperature remained freezing point.
The heater raised the temperature near freezing point.
- Look at the lines containing the phrase *the freezing point*. What words come after this phrase?
- Complete these sentences:
The freezing point of alcohol is than the freezing point of water.
..... will lower the freezing point of water.

Table 7.1 Data from a French–English parallel corpus of *Le Petit Prince/The Little Prince: on*

French	English
1. On en avale une par semaine et l'on n'éprouve plus le besoin de boire.	You need only swallow one pill a week, and you would feel no need of anything to drink.
2. Il faut s'astreindre régulièrement à arracher les baobabs dès qu'on les distingue d'avec les rosiers auxquels il ressemble beaucoup quand ils sont très jeunes.	You must see to it that you pull up regularly all the baobabs, at the very first moment when they can be distinguished from the rose-bushes which they resemble so closely in their earliest youth.
3. On épargne cinquante-trois minutes par semaine.	With these pills, you save fifty-three minutes in every week.
4. Quand on veut faire de l'esprit, il arrive que l'on mente un peu.	When one wishes to play the wit, he sometimes wanders a little from the truth.
5. Quand le mystère est trop impressionnant, on n'ose pas désobéir.	When a mystery is too overpowering, one dare not disobey.
6. Ils répètent ce qu'on leur dit . . .	They repeat whatever one says to them . . .
7. Donc, quand la moralité de l'explorateur paraît bonne, on fait une enquête sur sa découverte.	Then, when the moral character of the explorer is shown to be good, an inquiry is ordered into his discovery.
8. On note d'abord au crayon les récits des explorateurs.	The recitals of explorers are put down first in pencil.
9. On attend, pour noter à l'encre, que l'explorateur ait fourni des preuves.	One waits until the explorer has furnished proofs, before putting them down in ink.
10. On s'assoit sur une dune de sable. On ne voit rien.	One sits down on a desert sand dune, sees nothing, hears nothing.
11. On risque de pleurer un peu si l'on s'est laissé apprivoiser . . .	One runs the risk of weeping a little, if one lets himself be tamed . . .
12. C'est dur de se remettre au dessin, à mon âge, quand on n'a jamais fait d'autre tentatives que celle d'un boa fermé et celle d'un boa ouvert, à l'âge de six ans!	It is hard to take up drawing again at my age, when I have never made any pictures except those of the boa constrictor from the outside and the boa constrictor from the inside, since I was six.
13. On disait dans le livre : "Les serpents boas avalent leur proie tout entière, sans la mâcher."	In the book it said: "Boa constrictors swallow their prey whole, without chewing it."

Table 7.1 (continued)

French	English
14. C'est très utile, si l'on est égaré pendant la nuit.	If one gets lost in the night, such knowledge is valuable.
15. S'il s'agit d'une brindille de radis ou de rosier, on peut la laisser pousser comme elle veut.	If it is only a sprout of radish or the sprig of a rose-bush, one would let it grow wherever it might wish.
16. Voici mon secret. Il est très simple: on ne voit bien qu'avec le coeur.	"And now here is my secret, a very simple secret: It is only with the heart that one can see rightly;
17. Tu sais . . . quand on est tellement triste on aime les couchers de soleil . . .	"You know – one loves the sunset, when one is so sad . . ."
18. Tantôt je me dis: "On est distrait une fois ou l'autre, et ça suffit!"	But at another time I say to myself: "At some moment or other one is absent-minded, and that is enough."
19. Quand on a terminé sa toilette du matin, il faut faire soigneusement la toilette de la planète.	When you've finished your own toilet in the morning, then it is time to attend to the toilet of your planet, just so, with the greatest care.
20. Or un baobab, si l'on s'y-prend trop tard, on ne peut jamais plus s'en débarrasser.	A baobab is something you will never, never be able to get rid of if you attend to it too late.

(adapted from Rézeau – <http://www.uhb.fr/joseph.rezeau/concord.html>)

The integration of activities of this kind into a lesson and a syllabus will be discussed below.

Reciprocal learning and parallel concordances

Perhaps one of the most exciting innovations in language teaching of recent years is the development of reciprocal learning. Reciprocal learning occurs when two language learners are paired, each helping the other learn their language. For example, a French speaker learning English may be paired with an English speaker learning French. Parallel corpora may be used to aid reciprocal learning, and they are also useful for teaching translation or for more conventional language-learning in situations where all learners share a common first language.

As an example, Table 7.1 shows some sentences extracted from

parallel corpora in French and English, which have been identified by searching on the French pronoun *on*. This pronoun has several translation equivalents in English, and the examples in Table 7.1 are used to alert both English and French learners to this fact. (Note that these sentences have not been translated for the purposes of the exercise – they are the genuine translations made by the original translator.) This example comes from exercises written by Joseph Rézeau (<http://www.uhb.fr/joseph.rezeau/concord.html>; see Rézeau 2001 for more examples).

From the examples in Table 7.1 it can be seen that the French word *on* has been translated into English by *one* (4, 5, 6, 9, 10, 11, 14, 15, 16, 17, 18); *you* (1, 3, 19, 20); *I* (12); impersonal *it* (13); and by the passive (2, 7, 8). For the learner of French, this indicates that *on* has a wider range of uses than might be supposed. For the learner of English, it indicates that English has a range of expressions where French prefers the single word *on*.

A common objection to the use of parallel corpora is that the translations found are just that, translations, and that they may be less than felicitous. A present-day speaker of English may object to the frequent use of *one* in these examples, preferring perhaps *you* in example 11 or a paraphrase such as *Someone who is so sad loves sunsets* for example 17. Such examples can be discussed by the reciprocal learners. It is important that the translations are not assumed to be the only correct version, but that they indicate a range of possibilities.

Another example, this time from Tim Johns' web-site (<http://www.bham.ac.uk/johnstf>), is the French word *dont*, which I select because I personally find it extremely informative. A selection only is given in Table 7.2.

From these examples, learners are first asked simply to identify the various ways in which *dont* is expressed in English. Further tasks follow, focusing on the translations using *whose* and those using *of which*. The final task is to put missing words into French or English sentences. Here are a few examples:

(i) *Dans la salle des États, oeuvres de la Renaissance dont la Joconde de Léonardo de Vinci.*

In the Salle des États are Renaissance works, _____ the Mona Lisa by Leonardo da Vinci.

(ii) *Hérodote advoua son ignorance de la façon dont le nom d'Europe fut donnée au continent occidental.*

Herodotus admitted his ignorance of the way _____ which the name 'Europe' was given to the western continent.

Table 7.2: Data from a French–English parallel corpus: *dont*

French	English
'La Guerre des mondes' (1898) dont la libre adaptation sur les ondes par Orson Welles sema la panique aux Etats-Unis en 1938.	The War of the Worlds (1898), whose 1938 radio adaptation by Orson Welles created a wave of panic in the United States.
Inspiré par la pensée humaniste, le programme des décors, dont beaucoup ont été détruits, multipliait les allégories érudites, les références à la mythologie et à l'histoire antique ou contemporaine, ainsi que les louanges à la gloire du Roi.	Inspired by humanist thought, the allegorical decorations, many of which have been destroyed, represent mythology and ancient and modern history, as well as the glories of the king.
Traverser la cour de la Fontaine, dont l'eau pure était jadis réservée aux rois.	Cross the Fountain Courtyard, where the remarkably fresh water issuing from this fountain was reserved for the king.
De ce plan d'eau partent trois "rivières" dont le Rio Grande qui sépare l'Hotel Santa Fe et l'Hotel Cheyenne.	Three "rivers" run out of this lake, including the Rio Grande which separates the Hotel Santa Fe from the Hotel Cheyenne.
Dans la tradition française, les accouchements royaux étaient publics: dans cette chambre sont ainsi nés dix-neuf enfants de France, dont Louis XV et Philippe V d'Espagne.	In France, royal births were public events: in this room nineteen children of France were born, among them Louis XV and Philip V of Spain.

(adapted from Johns – <http://web.bham.ac.uk/johnstf/dont.htm>)

(iii) "*Nous sommes trop loin, effectivement,*" admit Ortiz, dont la voix _____.

"In fact we are too far off," Ortiz admitted, his voice trembling.

(iv) *Drogo regardait se profiler sur la poussière de la route l'ombre nette des deux chevaux dont les têtes, à chaque pas, faisaient "oui oui".*

Drogo watched the clear-cut shadows of the two horses on the dust of the road, their heads _____ at every step.

(v) *L'héritage dont il s'_____ est un héritage contesté . . .*
The heritage in question is a disputed heritage . . .

As I mentioned above, I regard reciprocal learning and the use of parallel corpora as two of the most exciting innovations in language

teaching and learning in recent years. Learners teaching each other are truly empowered, and are likely to be genuinely motivated to make discoveries about each other's language. The role of the teacher becomes that of materials-provider. My excitement about parallel concordances is a personal response to seeing English–French concordances, and finding from them enormous amounts of information about French. The use of *dont* before a noun phrase, translated as *including* or *among them*, and the use of *dont* + clause in translating 'their heads nodding' and so on were new to me. Moreover, this is not simply a matter of learning how to translate.² As a learner of French, having my attention drawn to these examples through their translations gives me a better 'feel' for how the word *dont* is used. The obvious restriction on reciprocal learning, however, is that it can be undertaken only in a context where there are students learning each other's language. This is not the situation in most contexts where English is being taught.

Corpora and language teaching methodology

Many teachers have two main reservations when considering the use of DDL in the classroom. The first is the means by which DDL can be integrated into the plan for an ordinary lesson. The second concerns the language points that seem to be the topic of DDL materials. These tend to deal with the minute details of the phraseology of particular words, and may be difficult to reconcile with the 'big themes' of language teaching, such as 'tenses' or 'articles'. In this section some answers to these questions are offered. The section draws largely on the work of Dave and Jane Willis in considering the place of corpus-based materials in language teaching.

DDL as consciousness-raising

DDL does not 'teach' a language feature, but presents learners with evidence and asks them to make hypotheses and draw conclusions. As an activity, it therefore fits best with a lesson that has such learner-centred activities built into it. An example of such an approach is the framework for task-based learning proposed by J. Willis (1996). Willis defines a task as 'a goal-oriented activity in which learners use language to achieve a real outcome' (1996: 53). She proposes a framework consisting of three stages (1996: 53):

² It is interesting that Johns' proposals re-visit traditional grammar-translation methods.

Pre-task: Introduction to the topic and task

Task cycle: Task → Planning → Report

Language Focus: Analysis and practice

For the Language Focus stage, Willis and Willis (1996) propose that ‘consciousness-raising activities’ should be used, designed to draw learners’ attention to some of the language features in the texts (written and spoken) that they have been engaged with when doing the task. Because a particular feature may occur only once or twice in the text, additional corpus material may be useful to help the learner to see a pattern rather than relying on a single occurrence.

As an example, here is one of the texts used as an illustration by Willis and Willis (1996: 71). It is a fairly simple (if frightening) story:

Auto-pilot

The flight ran several times a week taking holiday-makers to various resorts in the Mediterranean. On each flight, to reassure the passengers all was well, the captain would put the jet on to auto-pilot and he and all the crew would come aft into the cabin to greet the passengers.

Unfortunately on this particular flight the security door between the cabin and the flight deck jammed and left the captain and the crew stuck in the cabin. From that moment, in spite of efforts to open the door, the fate of the passengers and crew was sealed.

Willis and Willis suggest a variety of activities designed to encourage students to think about the lexis and grammar of the story. Below is an additional short sequence of activities, each one making use of corpus material in addition to the story itself.

The sequence begins by underlining the phrase: *left the captain and the crew stuck in the cabin*. This phrase from the story has been chosen because the verb *left* here occurs in a useful pattern, in which the verb is followed by a noun group and then by a past participle (*stuck*). Looking at the concordance lines for *LEAVE* in this pattern, it is striking that the past participles are usually words with a negative evaluation, such as *exhausted*, *crippled*, *paralyzed*, *shocked*. Here are some carefully selected examples:

The masked men left her bound and gagged.

A serious operation left her confined to a wheelchair.

A childhood illness has left her crippled . . .

The war left 300,000 homes destroyed.

The bitter winds left many anglers frozen to their seats.
An earthquake . . . killed around 170 people and left thousands homeless.

The subject may be a human being doing something intentionally (*the masked men*) but is more often an inanimate object without intentions (*a serious operation . . . a childhood illness*). The last example above uses a different but very similar pattern, in which the noun group is followed by an adjective instead of a past participle.

The teacher could draw attention to the pattern by asking learners to complete a table which isolates elements of the pattern, thus:

<i>The security door</i>	<i>left the captain and the crew</i>	<i>stuck in the cabin.</i>
<i>A serious operation</i>	<i>left her</i>	<i>confined to a wheelchair.</i>
<i>The bitter winds</i>	<i>left many anglers</i>	<i>frozen to their seats.</i>
<i>An earthquake</i>	<i>left thousands</i>	<i>homeless.</i>

This could be taken a step further by looking at other verbs with the same pattern. Here are some examples of *KEEP* and *FIND*:

Russian troops have kept the town sealed off since Saturday.
. . . a social life which kept us and others entertained . . .
I kept myself fit all summer.
She kept that world completely hidden from her friends.
He found himself immediately surrounded by opposing players.
Ray found himself charged with murder.
American soldiers found themselves hopelessly outnumbered.
At some point, he found himself drawn into conversation with Nina and her new friend.

Willis and Willis (1996: 66) refer to the 'grammar of class' as being an important component of pedagogic grammar. Here we are looking at a verb that belongs to a particular 'class': the class of verbs that are followed by a noun and a past participle or adjective.

Another important target of consciousness-raising mentioned by Willis and Willis is collocation. An example of a fixed collocation in the 'Auto-pilot' text is *fate . . . sealed*. This collocation is in danger of being missed by a reader of the text because the individual words are separated by a fairly long noun group: *of the passengers and crew*. (Lewis 1996: 14 makes the point that phenomena like this blur the 'word partnership'.) To make the word partnership, or collocation, clearer, the teacher might show these concordance lines:

These groups, who sealed the fate of President Marcos, have also lo
night appears to have sealed their fate. Buoyed up by the survival of t
Mr Wilson had sealed his fate shortly before the murder when he
nd that politics have sealed his fate – all these features being peculi

ner that should have sealed the fate of the tan once and for all, we
of the border that sealed Collins' fate. And the other directors
If ever Sir Richard's fate was sealed it was at that moment
to happen. But now my fate was sealed. In the morning the
rian regimes, and its fate was sealed for a time by the defe
1945, and in 1951 its fate was sealed, even though Labour
Alas, they cried, our fate is sealed. For the sake
She was convinced her fate was sealed and so she shut her eyes

Having been asked to identify the common pattern in these lines, the learners can be asked to find a similar example in the reading text.

The pedagogic corpus

In the examples given above, the concordance lines are taken from the Bank of English. D. Willis (1993, cited in Willis and Willis 1996: 67) suggests an alternative source: what he calls the learner's pedagogic corpus. This consists of all the language that a learner has been exposed to in the classroom – mainly the texts and exercises that the teacher has used. If the teacher has used authentic texts with a class, the corpus will consist of authentic language. If specially written texts have been used, the corpus will consist of invented language. The advantage of a pedagogic corpus is that, when an item is met in one text, examples from previous (and future) texts can be used as additional evidence for the learner to draw conclusions. The disadvantage of the general corpus – the unfamiliarity of the language in it – is overcome. Instead the teacher draws together for the learner aspects of the learner's past language experience to enable the learner to see patterns.

As an example, here are some concordance lines extracted from the first 50 pages of a Malaysian coursebook in English for post-elementary students (Khong et al 1987). The first set of lines is for the word *at*. These lines are taken from all parts of the course book, that is, from the rubrics to students as well as from the reading samples in the book.

I think we should leave it at the office.
Turn right at the junction.
ppens around 9:15am on Saturday mornings at the following places?
e lives at 23 Jalan Berenang. Norliza studied at the SRJK Jalan Cawang.
Ramlee's family will move into the house at No.1 Jalan Kiambang.
The school office is at the end of the corridor.
At the end of the corridor, turn left
t time does the afternoon session start? At ten minutes past one.
Doesn't the class begin at 9 am?
ident that happened one Saturday morning at around 9:15am.

She laughed as she looked at the timetable.
 notice board. Some students are looking at it excitedly.
 Look at these important benefits.
 Read the passage and look at the plan of the neighbourhood.
 How good are you at describing things. Let's find out
 I've one brother, no sisters at all.

By this point in the course book, the learners have met *at* in the following contexts:

- to indicate place;
- to indicate time;
- after *look*;
- after *good*;
- in the frame *no . . . at all*.

If a teacher wants to draw attention to the use of *at* in a new text, these concordance lines could be used as supplementary information.

The second example comes from the second 50 pages of the same book. Suppose the learners now meet an '-ing' form following *when* (e.g. *When reading this passage, . . .*). The teacher wishes to remind them of this use of *when* and similar words. Here are the examples from the course book:

Put back the newspapers after reading them.
 Look left, right and left again before crossing the road.
 He has also written down what one should not do while playing the game.
 When asking questions, ask only wh-questions.
 A student dropped her purse when getting into the school bus.

These examples are somewhat stilted because none of the texts in this course book are authentic. They might be supplemented with a few genuine examples:

When buying a chair, you should first consider its function and the price.
 If you suffer from headaches when reading . . .
 When buying clothes for your baby, I'd definitely go for convenience . . .
 Take care when using traditional remedies.
 Wear rubber gloves when washing up.

Corpora and syllabus design

In this chapter so far we have been taking the view-point of the classroom teacher and materials writer, and we have seen the contribution that materials based on concordances may make to the language class. There are, however, wider issues at stake. If, as Sinclair (1991: 100) says, 'language looks different when you look at

a lot of it at once', then the experience of using corpora should lead to rather different views of syllabus design. One type of syllabus whose design is based on concepts arising from corpus studies is the 'lexical syllabus'.

The notion of a 'lexical syllabus' was proposed in a paper by Sinclair and Renouf (1988), and finds its fullest exposition in D. Willis (1990). The term is occasionally (mis-)used to indicate a syllabus consisting only of vocabulary items, but as Sinclair, Renouf and Willis use the term, it comprises all aspects of language, differing from a conventional syllabus only in that the central concept of organisation is lexis. At its most simple, the argument is that it makes sense to teach the most frequent words in a language first. Sinclair and Renouf argue that 'the main focus of study should be on (a) the commonest word forms in the language; (b) the central patterns of usage; (c) the combinations which they usually form' (1988: 148). Their point is that the most frequent words have a variety of uses, so that learners acquire a flexibility of language fairly easily. In addition, the main uses of the most frequent words cover the main points of grammar, if in an unfamiliar form. Sinclair and Renouf quote *MAKE* as an example of a word with many uses, some of which are rarely covered in most beginners' courses. The most frequently occurring use of this verb is in combinations such as *make decisions*, *make discoveries*, *make arrangements*, rather than in the more concrete *make a cake* etc. In Sinclair's terminology, *MAKE* is used as a delexical verb more frequently than as an ordinary verb. An English course that focuses only on the concrete sense of *MAKE* denies the learner the opportunity to express sophisticated meanings with a simple verb.

Another example of a frequent word with multiple uses is *back*. This is a very frequent word: according to Sinclair 1999, it is 95th in frequency in the Bank of English, ahead of, for example, *get*, *may*, *how*, *think*, *even* and *us*. The reason for this frequency is that it is used in phrases such as *get the bus back*, *come/go back*, *look back*, *move back*, *turn back*, as well as as a noun: *behind your back*, *at the back*. Teaching the typical uses of *back* therefore introduces the learner to a large amount of language though not a massive vocabulary. Sinclair and Renouf make the point:

Almost paradoxically, the lexical syllabus does not encourage the piecemeal acquisition of a large vocabulary, especially initially. Instead, it concentrates on making full use of the words that the learner already has, at any particular stage. It teaches that there is far more general utility in the recombination of known elements than in the addition of less easily usable items. (1988: 155)

Turning to the issue of grammar in a lexical syllabus, Sinclair and Renouf argue that in a lexical syllabus, a separate listing of grammatical items is unnecessary:

If the analysis of the words and phrases has been done correctly, then all the relevant grammar etc should appear in a proper proportion. Verb tenses, for example, which are often the main organizing feature of a course, are combinations of some of the commonest words in the language. (1988: 155)

D. Willis (1990) takes up the issue of lexis and grammar (see also chapter 6). He points out that 'English is a lexical language', meaning that many of the concepts we traditionally think of as belonging to 'grammar' can be better handled as aspects of 'vocabulary'. For example, the passive can be seen as *BE* plus an adjective or past participle, rather than as a transformation of the active (1990: 17). Conditionals can be handled by looking at the hypothetical meaning of *would*, rather than by proposing a rule about sequence of tenses, that often does not work (1990: 18–19). He also argues that what is traditionally termed 'grammar' can often be called 'pattern' (1990:51). For example, a pattern consisting of 'noun phrase + *am/are/is* + . . . *ing*' is what is more usually called the present continuous tense. Other patterns that are less often treated as basic grammar might include other frequent words, such as *way*, e.g. '*the* + adjective + *way* + *of* + . . . *ing*' (*the best way of getting to Birmingham . . .*), '*the only way* + *that*-clause + *is/was* + *to*-infinitive clause' (*the only way you'll do that is to get the train*). In other words, Willis argues that the most productive way to interpret grammar in the classroom is as lexical patterning and, conversely, that all patterns involving frequent lexical items are important in the classroom, not only those that are traditionally covered by 'grammar'. Because patterns attach to all lexical items in the language, learning the lexis means learning the patterns and therefore the grammar.

Perhaps Willis' most radical suggestion is that a syllabus can, in effect, consist of a corpus (1990: 70). In other words, if the course designer collects pieces of authentic language that contain instances of the most frequent patterns of the most frequent words, then that collection (corpus) will exemplify what the learner needs to know. The job of the teacher or materials writer, then, is to devise ways of encouraging the learner to engage with the material in the corpus (e.g. by setting tasks) and of helping the learning to 'notice' (Schmidt 1990; Bernadini 2000) the patterning of language (e.g. by consciousness-raising activities). A description of the syllabus would, in effect, be a description of the corpus. If the syllabus was expressed as a list of items, it would be as a list of the most frequent word-forms in the

corpus, along with their most typical phraseologies. As the texts making up the corpus were presented to the learners, the syllabus would inevitably be covered. This alters the respective roles of the syllabus designer and materials writer quite considerably. Instead of the syllabus designer selecting items of language description and the materials writer choosing texts to illustrate them, the materials writer will choose interesting texts and the syllabus designer will keep a check on the balance of the overall collection of texts, ensuring that its most frequent word-forms, and their typical phraseologies, match what the learners require. Here, of course, there is an element of subjectivity. The syllabus designer may aim to mirror the distribution of structures, word frequency and phraseology in a larger, general corpus, or may decide that the learners' age, or specific needs, makes a different target corpus more appropriate. This subjectivity is no more than syllabus designers always employ, however, and has the advantage of making an appeal to principle, rather than to conventional wisdom. A syllabus of this kind would have the advantage of answering Long and Crookes' (1992: 33) objection that a lexical syllabus leads to artificial teaching materials if language is written specially to demonstrate key lexical items. Indeed, Willis' concept of a collection of texts is not dissimilar from the task-based syllabus proposed by Long and Crookes, though it would be a more concrete entity.

One problem in employing a corpus as syllabus is knowing how to describe the relevant frequencies in the corpus. A word-by-word account is very lengthy. One useful piece of supplementary information is a list of frequently occurring sequences. Sequences of this kind are of recent increasing interest to corpus linguists. De Cock et al (1998, also de Cock 1998) compare 'prefabs' in native-speaker and learner corpora, to test the hypothesis that learners tend not to use formulae as frequently as native speakers do. Biber et al (1999: 993–994) examine 'lexical bundles' in conversation and academic prose, using the Longman Grammar Corpus. They find that three-word bundles are much more frequent than four-word, that both kinds of bundles are more frequent in conversation than in academic prose, and that in conversation the bundles comprise more of the total word-count (28%) than they do in academic prose (20%). Some of the very frequent bundles in conversation include: *I don't know, I don't think, do you want, I don't want, don't want to, don't know what, and I said, I was going to, are you going to*; the frequent bundles in academic prose are, unsurprisingly, very different, and include: *in order to, one of the, part of the, the number of, the presence of, in the case of, on the other hand*.

D. Willis (1998) has done a similar study using the somewhat larger Bank of English corpus, but not differentiated by register. Some of the very frequent four-word combinations are:

Phrase	Number of occurrences	Phrase	Number of occurrences
the end of the	2,074	an awful lot of	514
a lot of people	1,834	in the middle of	510
nice to talk to	1,650	in the first place	477
that sort of thing	1,531	that kind of thing	441
a lot of the	1,189	this sort of thing	437
quite a lot of	1,098	per cent of the	392
a bit of a	1,089	got a lot of	389
end of the day	896	a little bit more	382
of the things that	654	a couple of years	366
the rest of the	608	a lot of time	351
a lot of money	595	a lot of things	346
a little bit of	570	most of the time	346
in terms of the	565	used to go to	337
to go to the	549	think a lot of	325
no no no no	536	to make sure that	324

The importance of these bundles or phrases is, firstly, that a syllabus designer working with a pedagogic corpus would wish to ensure that the corpus reflected these sorts of figures, if necessary differentiated by register, and secondly, that a materials designer would wish to draw attention to them as useful formulae for learners to use.

Challenges to the use of corpora in language teaching

Although corpora are widely acknowledged to be a valuable resource in describing language, there is less consensus on the value of corpus findings in the description of language for learners or on the use of corpus-based material in language classrooms. Among others, Widdowson (2000) and Cook (1998) have spoken against what they term an 'extreme' attitude towards using corpora in language teaching.³ At the risk of over-simplification, their arguments can be summarised thus:

³ For Widdowson, this is part of a more general argument against the uncritical use of theoretical linguistics in applied linguistics. Interestingly, Borsley and Ingham (forthcoming) regard corpora as the preoccupation of Applied Linguistics as opposed to theoretical linguistics.

1. A corpus is 'real language' only in a very limited sense. Language in a corpus is de-contextualised and must be re-contextualised in a pedagogic setting to make it real for learners. In Widdowson's (2000) terms, a corpus comprises traces of texts, not discourse.
2. Teachers (and course book writers etc) should not accept corpus evidence uncritically, but should appraise it in the light of other sources of information about language such as introspection and elicitation. In particular, frequency should not be the only factor in deciding what to teach: how salient a language feature is should also be taken into account, as should how highly valued a language item is. Learners should be encouraged to be creative in their language use, and should not be restricted to clichéd utterances.
3. Corpora tend to comprise the language of native speakers only, whereas many learners will never communicate with a native speaker and/or are not interested in native speaker norms. In particular, the details of phraseology or collocations may be unimportant to a non-native speaker of English. Too strong a dependence on corpora of native-speaker English tends to de-value the language of non-native-speakers and to perpetuate colonialist attitudes towards English.⁴
4. In a similar vein, learners should be allowed to approach language in a way they feel comfortable with. In many cases, this will be via grammatical rules and lists of lexical items. Learners should not be forced to approach English via 'lexical chunks' exclusively.

Some of these points can be taken as common ground. It would be very odd to suggest that language should not be contextualised within the classroom or that teachers should approach corpus evidence uncritically. Previous chapters in this book have stressed the need for caution in extrapolating from a corpus to a language and the importance of careful thought in interpreting corpus evidence. As Barlow (1996: 2) comments, 'using such powerful tools should not cause the researcher to become complacent and imagine that "language" is now in the computer'.

There are, however, three points here that deserve closer attention. These are: the issue of native-speaker corpora; the issue of frequency versus saliency, value and creativity; and the issue of lexis, grammar and 'lexical chunks'. Each of these will be dealt with in turn below.

I would not wish to argue against Cook's (1998) concern that

⁴ See, for example, Hall and Eggington (eds.) 2000.

corpora tend to treat native-speaker language as overly valuable.⁵ Cook seems to imply, however, that the English of non-native speakers (sometimes called International English, and exemplified by interactions between a Japanese manufacturer and a Turkish wholesaler) does not contain those features that corpus linguists claim for native-speaker English, such as variation between registers, restrictions on co-occurrence, association between pattern and meaning, and so on. Hunston and Francis (1999: 268–270) suggest that although patterning in International English might be different from that found in any native-speaker variety, it would still exist and be worth teaching as patterning. Their argument is that the process of ‘doing corpus research’ has a value that is independent of the value of the product on which that research is currently carried out. If currently available corpora are inadequate, and in this respect they very clearly are, then there is a strong argument for compiling more adequate corpora, in this case of International English, rather than simply abandoning corpora altogether. I suspect that compiling such a corpus would be fraught with difficulties, ranging from ‘Who would consider such an enterprise worth funding?’ to ‘Whose language should be collected?’, but the very existence of such questions, and the problems they raise, is itself usefully revealing of attitudes towards International English.

The second interesting question that Cook raises is that of the importance of frequency. It is very commonly argued by those who advocate using corpus evidence in teaching that what is most frequent should be taught first, and that learners’ attention should be drawn most to frequently occurring phenomena. The opposing argument is that certain aspects of English are important even though they are not frequent, either because they carry a lot of information or because they have a resonance for a cultural group or even for an individual. Wray and Perkins (2000, and citing Hickey 1993 and Howarth 1998) make a similar argument when they suggest that a sequence of words may constitute a ‘formula’ for an individual or a cultural group, even if the sequence is attested only rarely. Items which are important though infrequent seem to be those that echo texts which have a high cultural value. A good example is the co-occurrence of *death* and *adventure* in the following extract from J. K. Rowling’s *Harry Potter and the Philosopher’s Stone*:

⁵ It is worth adding, perhaps, that this argument applies only to English, because of its unique hegemony in the modern world. Learners of French, German, Japanese etc might be less disconcerted by having access to a corpus of language produced by native speakers only.

Death is but the next great adventure.

For many (though not all) readers, *death* and *adventure* may effectively be collocates, because of an intertextual reference to the classic children's play *Peter Pan*, by J. M. Barrie, which includes the line:

To die will be an awfully big adventure.

There are a handful of similar instances in the Bank of English, some specifically quoting Barrie, but too few to have statistical significance. The resonance of Rowling's phrase, then, apparently comes from its cultural salience, not its frequency.⁶

In many cases, however, cultural salience is not so clearly at odds with frequency. Examples of salient items sometimes given are: proverbs such as 'Too many cooks spoil the broth'; slogans such as the American Express 'That will do nicely', which has a resonance arising from the frequency with which the advertisement was repeated and mimicked; and headlines such as 'Gotcha!' (the *Sun* newspaper's infamous response to the sinking of the Argentinian ship the *Belgrano* prior to the Malvinas/Falklands conflict in 1982). In each case, salience does seem to be reflected in statistical measures. *Spoil* is a significant collocate of *broth* (occurring two places to the left of *broth* with a high MI-score of 12, indicating a strong collocation, and a t-score of 3.3, indicating a certainty just above the cut-off point of 2), and the proverb is not only quoted but exploited with variation, as these concordance lines illustrate:

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

Similarly, *do* occurring immediately to the left of *nicely* has a t-score significance of 10.7, with most examples clearly echoing the advertising slogan, either directly or through exploitations such as *tatt will do nicely*, *data'll do nicely* or *American Express won't do nicely*. *Gotcha* occurs in the Bank of English 128 times, including in the phrase *gotcha journalism*.

In these examples, the corpus examples do not explain why a phrase is significant, but the frequency information does seem to follow the salience. The fact that salient phrases are often subject to variation is also illustrated, in turn showing that English is about creativity as well as cliché.

⁶ I am grateful to Hsin Chin Lee for bringing this example to my attention.

Another aspect of saliency is discussed by Barlow (1996) and Shortall (1999), though not using that terminology. Barlow suggests that learners create schemata for grammatical features of a language, contrasting this with the 'parameter setting' hypothesis. These schemata are based partly on the evidence that the learner meets when experiencing authentic examples of the language, but also on the prototypes or expectations that the learner has about what meaning distinctions might be made. Arguing that both induction and expectation have a role in language learning, Barlow comments: 'the learner is not seen as just a passive pattern extractor, but is, in addition, a cognizer with the ability to make numerous cognitive distinctions' (1996: 17–18). Shortall goes further in relating this to teaching syllabuses. He points out that all language users have 'prototypes' about aspects of language use, and that these may conflict with the evidence of what is most frequent. For example, he finds that, when asked to produce a sentence with *there*, most people use a concrete noun and a prepositional phrase, as in *there are three books on the table*, whereas in the Bank of English corpus *there* constructions are more frequently used with abstract nouns and clauses, as in *there is evidence to suggest that . . .* Shortall expresses the teacher's resulting dilemma in the form of two conflicting statements: 'If concrete nouns are prototypical, and if this is the kind of noun people first think of, perhaps these should be taught first in EFL textbooks' and 'If abstract nouns are more frequent in real language (or in the corpus) perhaps these should be given priority.' He argues that prototypes are so strong that learners should be taught them first and only later introduced to the more frequent usages. In this, Shortall demonstrates the discerning attitude towards corpus evidence that Widdowson and Cook advocate.

The third point I wish to debate here is Cook's observation that learners should not be forced to restrict their learning experience to 'lexical chunks'. If a learner wishes to perceive English in terms of grammatical rules supplemented by vocabulary lists, she or he should be allowed to do so, and not be forced into ignoring rules and learning only phrases. If researchers into corpora did advocate a 'phrase-book' approach to language learning, then Cook's criticism would be legitimate, but this is far from the case. The essence of the 'idiom principle' and of 'units of meaning', as discussed in chapter 6, is that the patterning of language is more flexible and also more pervasive than the concept of 'lexical chunks' would suggest. Again, Barlow (1996: 15) expresses this well:

Part of the motivation for [this] approach . . . is a rejection of the

distinction between a creative, compositional, productive component of the grammar and a component consisting of a collection of fixed idiomatic forms. The claim is that *most* of language consists of semi-regular, semi-fixed phrases or units . . . [words] have an affinity for each other and are linked together, but not so strongly as to form an identifiable lexical unit.

Barlow's discussion of the use of reflexive pronouns illustrates this phenomenon. Reflexive pronouns are used predominantly with some verbs rather than others (see Francis et al 1996 for detailed lists), but these do not constitute 'fixed phrases'.

Another point to be made here is that where phrases are advocated as a useful input to language learning (see, for example, the discussion of Willis above), the notion of language teaching is somewhat different to that apparently envisaged by Cook. It is not recommended that teachers 'present' phrases as a teaching item, but that phrases are among the variety of lexical and grammatical features which are amenable to consciousness-raising (D. Willis, personal communication). Thus, the learner's predilection for viewing language in a particular way is not thwarted, but may be encouraged to expand.

Widdowson argues that the importance of corpora 'lies not in the answers they provide but in the questions they provoke' (2000: 23). I would agree with the words here, though not in the way Widdowson means them: possibly the most far-reaching influence of corpora is not the individual observations that have been made using them, but the radical questions they have raised about the nature of language itself (see chapter 6). One of the questions provoked for Widdowson is 'If they do not represent real language for the learner, then what *does*?' (2000: 23). Given the ambiguity of this question, it is one that we can all probably agree is worth answering.

Conclusion

In this chapter, general issues concerning the way that corpora can inform classroom teaching have been discussed, as they affect the syllabus designer, the materials writer and the teacher. In the next chapter, more specific issues relating to language teaching will be taken up, focusing on English for Academic Purposes and language testing. The use of learner corpora will also be demonstrated.