Unit A7
Corpus analysis and academic texts

Genre descriptions have to be based on sufficient text samples to ensure that the principles and regularities observed are representative of the target genre, and genre analysts have been greatly assisted in this by the use of language corpora. A *corpus* is a collection of naturally occurring texts used for linguistic study. While a corpus does not contain any new theories about language, it can offer fresh insights on familiar, but perhaps unnoticed, features of language use. This is because a corpus is a more reliable guide to what language is like than human intuition. While we all have experience of certain genres, much of this remains hidden, so that, for example, even the best teachers are often unable to explain to their students why some phrasing or expression is preferred over another in a given context. A corpus, in other words, provides an evidence-based approach to language teaching.

The idea behind a corpus is that it represents a speaker's experience of language in some domain. This makes the approach ideal for studying the features of academic genres as it means we can describe them more accurately so students can learn to use them more effectively. Using any one of a number of commercially available, and relatively inexpensive, text analysis programmes (*concordancers*), teachers can selectively examine fairly large amounts of texts to supplement their intuitions, not to confirm whether something is possible or not, but to describe whether it is frequent or not. As Sinclair (1991: 17) points out, this moves the study of language away from ideas of what is correct, towards what is typical or frequent.

Task A7.1

➤ How might a corpus be of value to you as an EAP teacher? What kinds of texts would be most useful to your students? How would you use a corpus in your course?

CORPUS STUDIES AND FREQUENCY

The idea of *frequency* is central to corpus studies as corpora are not concerned with what can occur in a genre or register but with what frequently and typically occurs. In other words, priority is given to describing the commonest uses of the commonest words on the assumption that if something is observed to happen often enough in the past then it is likely to be significant in the future too. This allows us to predict the ways that other representative examples of the genre will be organized and the features it is likely to contain. Corpus analyses therefore often begin by automatically counting the frequency of words or grammatical patterns in order to characterize the domain under study.

Corpus studies have shown that the most frequent words in English cover an inordinate percentage of any text, with the top three words (*the, of, to*) making up some 10 per cent of the 400 million words in the Bank of English corpus, for instance, and the first 100 comprising about one-half of all written and spoken texts (e.g. Hunston, 2002). The most frequent words in any corpus are therefore grammatical words, but working down frequency lists soon reveals key items in that genre, enabling the teacher to identify and teach basic items in their classes. Table A7.1 shows differences in undergraduate course books in two disciplines.

Table A7.1 Most frequent nouns in introductory textbooks in two disciplines

Applied linguistics			*Biology*		
No.	*% of total*	*Word*	*No.*	*% of total*	*Word*
423	0.8663	language	166	0.4304	species
149	0.3052	speech	150	0.3889	DNA
128	0.2622	example	143	0.3708	spores
127	0.2601	interaction	135	0.3500	organisms
106	0.2171	act	117	0.3033	bacteria
101	0.2069	communication	116	0.3008	fungi
97	0.1987	students	95	0.2463	figure
93	0.1905	text	89	0.2307	organism
93	0.1905	acquisition	75	0.1945	RNA
91	0.1864	acts	68	0.1763	spore
90	0.1843	face	62	0.1607	cells
89	0.1823	input	59	0.1530	section
86	0.1761	rules	58	0.1504	genus
85	0.1741	communicative	55	0.1426	cell
79	0.1618	knowledge	49	0.1270	disease

One use of frequency counts in EAP is the construction of vocabulary lists such as the Academic Word List (Coxhead, 2000). These are based on the idea that vocabulary falls into three main groups (Nation, 2001):

- High-frequency words such as those included in West's (1953) General Service List of the most widely useful 2,000 word families in English, which provides coverage of about 80 per cent of most texts.
- An academic vocabulary of words which are reasonably frequent in academic writing across disciplines and genres and comprise some 8 per cent to 10 per cent of running words of academic texts.

- A technical vocabulary which differs by subject area and covers up to 5 per cent of texts.

Students are said to find such an academic vocabulary a particularly challenging aspect of their learning (Li and Pemberton, 1994). This is because, while technical vocabulary is central to the students' specialized areas, general academic words serve a largely supportive role and are 'not likely to be glossed by the content teacher' (Flowerdew, 1993: 236). But while general academic word lists are useful for EAP materials developers, we need to be cautious about them. It remains unclear how far a single inventory can represent the vocabulary of 'academic discourse', or how far it might be useful to students irrespective of their field of study (Hyland and Tse, 2007). Individual items tend to have different frequencies and meanings in different disciplines and genres, encouraging us to look beyond common core features and the autonomous views of literacy that such lists assume to recognize that contextual factors are crucial to language choices.

More sophisticated information can be gathered using software which counts not only words, but also grammatical features. By a semi-automatic procedure known as *tagging*, codes can be added to each word indicating its part of speech, so, for instance, the word *research* is tagged as either a noun or a verb each time it occurs, allowing much more detailed analyses of target genres. Biber's (1988) research, for instance, shows how written academic prose is characterized by bundles of grammatical features such as frequent nouns, long words, attributive adjectives and prepositional phrases which function to present densely packed information. In contrast, second-person pronouns, direct questions, present-tense verbs, private verbs (*feel, think*) and *that* deletions are less frequent because of their more interactive character. A tagged corpus can assist teachers in deciding on the relative merits of recommending past or present tense when teaching report genres, for example, or whether it is more useful to focus on active or passive constructions in essay writing.

Frequency counts are also a useful way of determining the features which are overused or under-used in the writing of L2 students in given genres. Research by Granger (1998) and Hinkel (2002) on learner corpora, for instance, shows that L2 academic essays contain a smaller range of vocabulary than L1 essays and are characterized by stylistic features more typical of informal speech than written discourse. A good example of a learner corpus informing classroom practice is Milton's (1999) study of his students' use of fixed expressions in their essays (e.g. Nattinger and De Carrico, 1998). Lacking good models of target academic genres, they seemed to fall back on a limited number of prefabricated 'lexical bundles' to avoid grammatical errors, leading them to a repetitive style of writing. By comparing a student essay corpus with a parallel corpus of L1 essays, Hong Kong school textbooks and published research articles, Milton confirmed that the L2 students used the same phrases far more often than L1 writers and was able to compile a list of alternative phrases from the L1 samples which he then included in his classes to help his students vary their academic writing (Table A7.2).

Table A7.2 Phrases in a Hong Kong learner corpus

	Frequency of phrases per 50,000 words in each corpus			
Lexical phrases with greatest difference	*L2 student texts*	*L1 student texts*	*School textbooks*	*Published articles*
Not used in L2 student texts				
In the/this case	0	9	11	16
It has also been	0	8	0	5
It can be seen that	0	8	0	4
An example of this is	0	8	0	3
This is not to say that	0	7	0	2
Overused in L2 student texts				
First of all	170	1	13	5
On the other hand	239	31	25	30
(As) we/you know	118	2	22	3
In my opinion	110	12	8	0
All in all	59	2	1	0

Source: Milton (1999: 226).

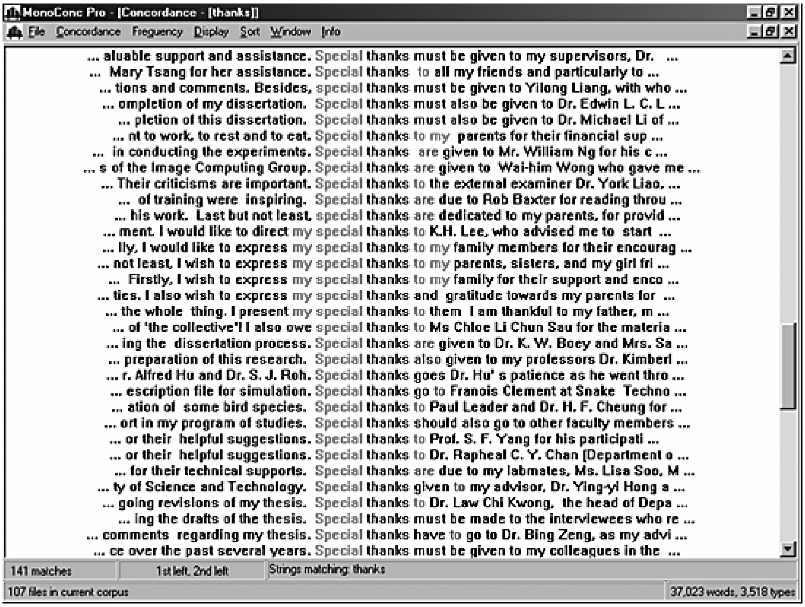
Task A7.2

➤ Why might frequency counts be useful to analysts or teachers? Do you think it would be more useful for students to discover word or pattern frequencies for themselves or to be given this information by teachers?

CONCORDANCING

In addition to frequency counts, analysts also explore corpora by examining concordances. A concordance brings together all instances of a search word or phrase in the corpus as a list of unconnected lines of text with the node word in the centre together with a sample of its linguistic environments. These lines therefore give instances of language *use* when read horizontally and evidence of *system* when read vertically. This makes it possible for the user to see regularities in its use that might otherwise be missed.

Moreover, by sorting the concordance lines by the first word to the left or to the right of the search word, frequent co-occurrences become visible. Thus in the study of dissertation acknowledgements mentioned earlier we discovered a strong tendency to use the noun *thanks* in preference to other expressions of gratitude (Hyland and Tse, 2004). By sorting concordance lines on the word to the left of this search word, we then found that this noun was modified by only three adjectives: *special*, *sincere* and *deep*, with *special* making up over two-thirds of all cases. Figure A7.1 is a screenshot from the program MonoConc Pro showing part of the results of this sorting.



```
... aluable support and assistance. Special thanks must be given to my supervisors, Dr.   ...
...  Mary Tsang for her assistance. Special thanks  to all my friends and particularly to ...
  ... tions and comments. Besides, special thanks must be given to Yilong Liang, with who ...
   ... ompletion of my dissertation.  Special thanks must also be given to Dr. Edwin L. C. L ...
      ... pletion of this dissertation.  Special thanks must also be given to Dr. Michael Li of ...
     ... nt to work, to rest and to eat. Special thanks to my  parents for their financial sup ...
  ...  in conducting the experiments. Special thanks  are given to Mr. William Ng for his c ...
... s of the Image Computing Group. Special thanks are given to  Wai-him Wong who gave me ...
   ... Their criticisms are important. Special thanks to the external examiner Dr. York Liao. ...
      ...  of training were  inspiring.  Special thanks are due to Rob Baxter for reading throu ...
     ... his work.  Last but not least, special thanks are dedicated to my parents, for provid ...
   ... ment. I would like to direct my special thanks to K.H. Lee, who advised me to  start  ...
   ... lly, I would like to express my special thanks to my family members for their encourag ...
  ... not least, I wish to express my special thanks to my parents, sisters, and my girl fri ...
    ...  Firstly, I wish to express my special thanks to my family for their support and enco ...
  ... ties. I also wish to express my special thanks and  gratitude towards my parents for  ...
   ... the whole  thing. I present my special thanks to them  I am thankful to my father, m ...
     ... of 'the collective'! I also owe special thanks to Ms Chloe Li Chun Sau for the materia ...
    ... ing the  dissertation process. Special thanks are given to Dr. K. W. Boey and Mrs. Sa ...
    ... preparation of this research.  Special thanks also given to my professors Dr. Kimberl ...
    ... r. Alfred Hu and Dr. S. J. Roh. Special thanks goes Dr. Hu' s patience as he went thro ...
     ... escription file for simulation. Special thanks go to Franois Clement at Snake  Techno ...
     ... ation of  some bird species.  Special thanks to Paul Leader and Dr. H. F. Cheung for ...
    ... ort in my program of studies.  Special thanks should also go to other faculty members ...
    ... or their  helpful suggestions. Special thanks to Prof. S. F. Yang for his participati ...
    ... or their  helpful suggestions. Special thanks to Dr. Rapheal C. Y. Chan (Department o ...
     ... for their technical supports.  Special thanks are due to my labmates, Ms. Lisa Soo, M ...
  ... ty of Science and Technology.  Special thanks given to my advisor, Dr. Ying-yi Hong a ...
   ... going revisions of my thesis.  Special thanks to Dr. Law Chi Kwong,  the head of Depa ...
      ... ing the drafts of the thesis.  Special thanks must be made to the interviewees who re ...
 ... comments  regarding my thesis. Special thanks have to go to Dr. Bing Zeng, as my advi ...
  ... ce over the past several years. Special thanks must be given to my colleagues in the  ...
141 matches        1st left, 2nd left        Strings matching: thanks
107 files in current corpus                                        37,023 words, 3,518 types
```

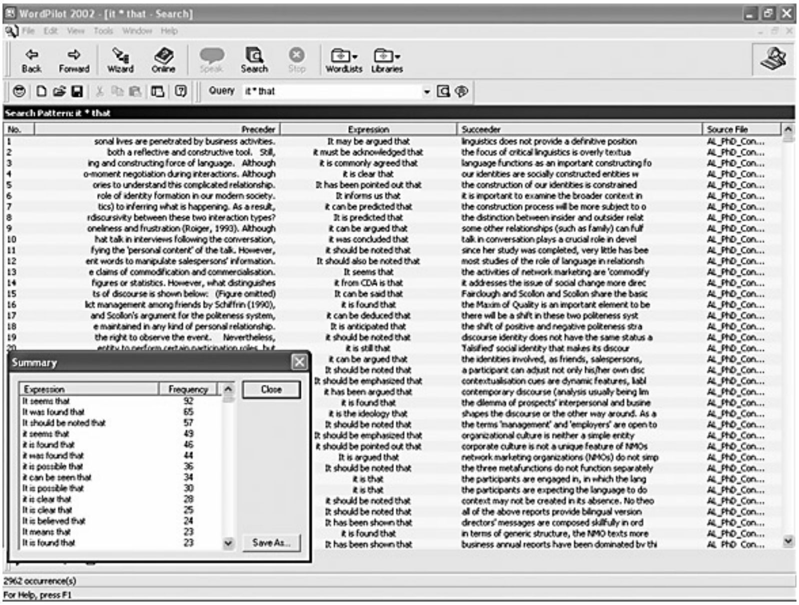
Figure A7.1 Screenshot from MonoConc Pro, showing a left sort on the word 'thanks'

Concordancing also allows searches for word combinations, even revealing frequencies and meanings of key phrases which vary by intervening words. Thus using the * wild card by entering the expression *it* * *that* will search for the word *it* followed by *that* in the near vicinity, producing examples such as these in a corpus of abstracts in research papers:

it is likely that
it seems that
it is clear that
it shows that
it is claimed that
it is true that
it is worth noting that
it is shown that
it is more likely that

When these examples are studied more carefully, they show that academic writers use this phrasing extremely frequently to express their evaluation of whether the following statement is likely to be true or not. In addition, the results show that expressions of certainty occur more often than those expressing doubt. This kind of information can help student writers not only to make use of this collocation in their own writing, but to use it in effective ways. Figure A7.2 shows a screenshot of concordance lines for this structure using WordPilot 2002 with a pop-up window listing the most frequent collocations.

The analysis of potentially productive phrases such as this is particularly useful for helping student writers to see how high-frequency grammar words often occur in



Query: it * that

Search Pattern: it * that

No.	Preceder	Expression	Succeeder	Source File
1	sonal lives are penetrated by business activities.	It may be argued that	linguistics does not provide a definitive position	AL_PhD_Con...
2	both a reflective and constructive tool. Still,	it must be acknowledged that	the focus of critical linguistics is overly textua	AL_PhD_Con...
3	ing and constructing force of language. Although	it is commonly agreed that	language functions as an important constructing fo	AL_PhD_Con...
4	o-moment negotiation during interactions. Although	it is clear that	our identities are socially constructed entities w	AL_PhD_Con...
5	ories to understand this complicated relationship.	It has been pointed out that	the construction of our identities is constrained	AL_PhD_Con...
6	role of identity formation in our modern society.	It informs us that	it is important to examine the broader context in	AL_PhD_Con...
7	tics) to inferring what is happening. As a result,	it can be predicted that	the construction process will be more subject to o	AL_PhD_Con...
8	rdiscursivity between these two interaction types?	It is predicted that	the distinction between insider and outsider relat	AL_PhD_Con...
9	oneliness and frustration (Roiger, 1993). Although	it can be argued that	some other relationships (such as family) can fulf	AL_PhD_Con...
10	hat talk in interviews following the conversation,	it was concluded that	talk in conversation plays a crucial role in devel	AL_PhD_Con...
11	fying the 'personal content' of the talk. However,	it should be noted that	since her study was completed, very little has bee	AL_PhD_Con...
12	ent words to manipulate salespersons' information.	It should also be noted that	most studies of the role of language in relationsh	AL_PhD_Con...
13	e claims of commodification and commercialisation.	It seems that	the activities of network marketing are 'commodify	AL_PhD_Con...
14	figures or statistics. However, what distinguishes	it from CDA is that	it addresses the issue of social change more direc	AL_PhD_Con...
15	ts of discourse is shown below: (Figure omitted)	It can be said that	Fairclough and Scollon and Scollon share the basic	AL_PhD_Con...
16	lict management among friends by Schiffrin (1990),	it is found that	the Maxim of Quality is an important element to be	AL_PhD_Con...
17	and Scollon's argument for the politeness system,	it can be deduced that	there will be a shift in these two politeness syst	AL_PhD_Con...
18	e maintained in any kind of personal relationship.	It is anticipated that	the shift of positive and negative politeness stra	AL_PhD_Con...
19	the right to observe the event. Nevertheless,	it should be noted that	discourse identity does not have the same status a	AL_PhD_Con...
20	entity to perform certain participation roles, but	it is still that	'falsified' social identity that makes its discour	AL_PhD_Con...
		it can be argued that	the identities involved, as friends, salespersons,	AL_PhD_Con...
		It should be noted that	a participant can adjust not only his/her own disc	AL_PhD_Con...
		It should be emphasized that	contextualisation cues are dynamic features, labl	AL_PhD_Con...
		it has been argued that	contemporary discourse (analysis usually being lim	AL_PhD_Con...
		it is found that	the dilemma of prospects' interpersonal and busine	AL_PhD_Con...
		it is the ideology that	shapes the discourse or the other way around. As a	AL_PhD_Con...
		It should be noted that	the terms 'management' and 'employers' are open to	AL_PhD_Con...
		It should be emphasized that	organizational culture is neither a simple entity	AL_PhD_Con...
		it should be pointed out that	corporate culture is not a unique feature of NMOs	AL_PhD_Con...
		It is argued that	network marketing organizations (NMOs) do not simp	AL_PhD_Con...
		It should be noted that	the three metafunctions do not function separately	AL_PhD_Con...
		it is that	the participants are engaged in, in which the lang	AL_PhD_Con...
		it is that	the participants are expecting the language to do	AL_PhD_Con...
		it should be noted that	context may not be created in its absence. No theo	AL_PhD_Con...
		It should be noted that	all of the above reports provide bilingual version	AL_PhD_Con...
		It has been shown that	directors' messages are composed skilfully in ord	AL_PhD_Con...
		it is found that	in terms of generic structure, the NMO texts more	AL_PhD_Con...
		It has been shown that	business annual reports have been dominated by thi	AL_PhD_Con...

Expression	Frequency
It seems that	92
It was found that	65
It should be noted that	57
it seems that	49
it is found that	46
it was found that	44
it is possible that	36
it can be seen that	34
It is possible that	30
it is clear that	28
It is clear that	25
It is believed that	24
It means that	23
It is found that	23

2962 occurrence(s)

For Help, press F1

Figure A7.2 Screenshot from WordPilot, showing concordance for 'it * that' in dissertations

regular patterns, even though the lexical items within these patterns may be less frequent. Armed with this kind of information about their target genres, EAP students are able to make choices which are better informed, guided by 'expert' practice and disciplinary expectations.

In addition, corpus evidence offers a range of information for EAP teachers and learners. For instance, collocation patterns can reveal features such as the following:

- The patterns of various forms, e.g. whether first-person pronouns are associated with claims, criticisms or research procedures in academic research papers.
- The differences between words which students often confuse, e.g. *bored* versus *boring, interested* versus *interesting, possible to* versus *possible that*, etc.
- The most appropriate words to use – e.g. whether to use the preposition *in, that* or *to* with *interested* and *interesting.*
- 'Semantic prosody', or the connotative meanings a word acquires because of its regular association with other words, e.g. the word *commit* carries unfavourable implications because of its regular co-occurrence with words such as *crime, murder, mistakes*, etc. Similarly the word *rife* has unfavourable semantic prosody (Partington, 1998: 67).
- Stable lexical patterning in particular disciplines, particularly nominal groups, e.g. *critical discourse analysis* or *static electric field.*

- The specific meanings that words take on in particular disciplines, e.g. *wall, energy, structure, concentration, body*, etc., in biology.
- How words change their meaning as a result of the surrounding text, e.g. the word *quite* boosts the meaning of non-gradable words such as *impossible, definitely* and *agree*, and hedges gradable words such as *interesting, beautiful* and *cynical.*

To summarize, the computer analysis of text corpora is an invaluable tool for EAP teachers. It indicates the high-frequency words, phrases and grammatical structures which characterize a given genre or discipline and reveals how these are typically used in patterns of collocation, or association, with other words or phrases. This, in turn, can help teachers to better understand the texts they teach and students to become more aware of the options available to them when communicating in their disciplines.

Task A7.3

➤ How could such concordances be built into learner exercises and tasks? Think of a task you could give to a group of students using a corpus. What problems might students have with concordancing as a classroom tool and how might you overcome those problems?