hallenge to
seeking to
'ten treated
and advice
know more
ature. They
kely to find

ce as it is a
municating
l of appeal-
s commun-
voice, but
s the writer
s can critic-
l accede to

as been no
which par-
ne effect of
estricted to
at students
eports, and
rly straight-
choices and
and genres.
ion to spe-
sity writing
ıtly require
remain, this
ting and to

Chapter 7

Corpora and corpus analysis: new windows on academic writing

Christopher Tribble

INTRODUCTION

The purpose of this chapter is to outline ways in which appropriate corpus resources may help students to develop competence as writers within specific academic domains. I have discussed elsewhere (Tribble, 1997a) the sets of knowledge which writers (in general) need in order to produce appropriate and effective texts. These can be summarised as in Table 7.1.

Such a framework brings together what are commonly called *process approaches* to writing instruction (Emig, 1983; Graves, 1984; Grabe and Kaplan, 1996) and *genre approaches* (Bhatia, 1993; Cope and Kalantzis, 1993; Halliday, 1978; Johns, 1994; Martin, 1989; Swales, 1990), and provides the basis for an integrated writing instruction programme. Thus, in the case of EAP students in higher education in Britain, learners (as a result of the academic programmes they are following) will most often be the ones responsible for the content knowledge that will be realised in writing activities. Teachers, on the other hand, will have the responsibility for creating opportunities in which learners can come to a fuller understanding of (a) the processes that are necessary to the completion of a writing task, (b) the institutional and contextual

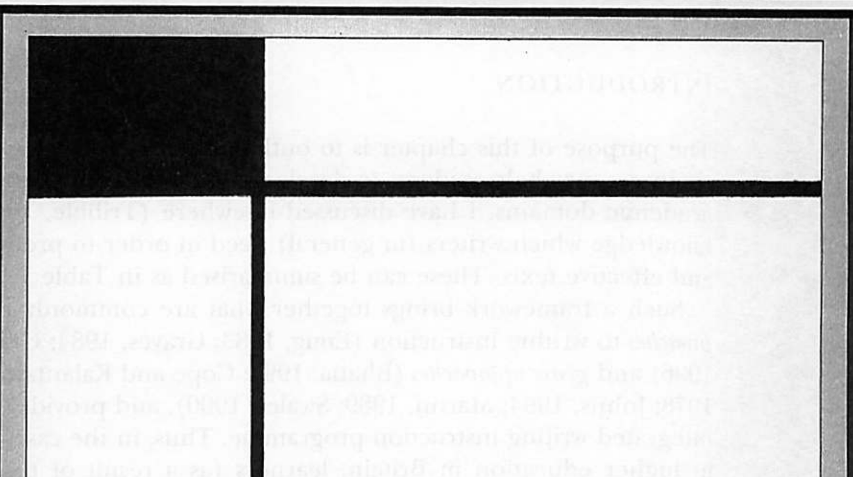
Table 7.1 What writers need to know

content knowledge	knowledge of the concepts involved in the subject area
writing process knowledge	knowledge of the most appropriate way of carrying out a specific writing task
context knowledge	knowledge of the social context in which the text will be read, and co-texts related to the writing task in hand
language system knowledge	knowledge of those aspects of the language system necessary for the completion of the task

Source: Based on Tribble, 1997a: 43

ACADEMIC DISCOURSE

Edited by John Flowerdew



guistics and Language Study

ditor: C. N. Candlin

constraints which operate in the target environment and determine what constitutes an allowable contribution, and (c) the linguistic choices which have to be made in order to produce such allowable contributions.

The question then arises as to how teachers can help learners to develop an understanding of the way texts work in social contexts, and how language use and communicative purpose intersect in the genres in which they have an interest. Language corpora may be one of the resources that teachers and learners will want to draw on.

Although this chapter is about ways in which the computer analysis of collections of academic texts can increase our understanding of academic writing, I will begin my discussion by looking at a single example: a short published paper in the Reading Academic Text Corpus (RAT).[1] (See the Appendix for the complete text.) As the text was written by a highly regarded scholar at the invitation of the editors of 'Plant Molecular Biology Reporter' (published by the Plant Molecular Biology Association), we can assume that it is felt to be 'situationally effective' (Bazerman, 1994a: 23) by members of a particular discourse community. In other words, it is the result of what Bazerman calls an 'expert performance' (Bazerman, 1994a: 131), and of potential interest to apprentice writers in that field.

In the second part of the chapter, based on this example analysis, I will make suggestions for what I will call a corpus informed approach to EAP writing instruction. I am working this way round – from micro to macro – because this analysis will enable me to demonstrate some of the ways in which corpus linguistic tools and resources, combined with discourse and genre analytic frameworks for analysis, can contribute to EAP learning and teaching.

USING CORPORA

To date, corpus linguistics has largely been driven by the needs of lexicographers, descriptive linguists, and the NLP research community (Aston, 1995; Tribble, 1997b; L. Flowerdew, this volume). This has created an overall push towards a 'biggest is best' view of the corpus (e.g. Sinclair, 1991), which, while it may be valid for these communities, does not necessarily meet the needs of teachers and learners in English for Academic Purposes programmes. The large corpus, whether it is 'balanced' as in the case of the British National Corpus (Burnard, 1995), or a monitor corpus as with the Bank of English at Birmingham University (Sinclair, 1991), provides either too much data across too large a spectrum, or too little focused data, to be directly helpful to learners with specific learning purposes. Although they have made, and will continue to make, an invaluable contribution to ELT lexicography and language description, large corpora appear to have less relevance to EAP writing instruction and other areas of ELT. It is for this reason that I suggest that using small corpus resources *alongside* bigger corpora can be helpful in

Table 7.2 Analytic framework (Contextual)

CONTEXTUAL Analysis	
1. name	What is the name of the genre of which this text is an exemplar?
2. social context	In what social setting is this kind of text typically produced? What constraints and obligations does this setting impose on writers and readers?
3. communicative purpose	What is the communicative purpose of this text?
4. roles	What roles may be required of writers and readers in this genre?
5. cultural values	What shared cultural values may be required of writers and readers in this genre?
6. text context	What knowledge of other texts may be required of writers and readers in this genre?
7. formal text features	What shared knowledge of formal text features (conventions) is required to write effectively into this genre?
LINGUISTIC Analysis	
8. lexico-grammatical features	What lexico-grammatical features of the text are statistically prominent and stylistically salient?
9. text relations/textual patterning	Can textual patterns be identified in the text? What is the reason for such textual patterning?
10. text structure	How is the text organised as a series of units of meaning? What is the reason for this organisation?

developing an understanding of academic written discourse. Such an approach fits in with current work that focuses on the value of smaller corpora for language learning and teaching (see J. Flowerdew, 1993a; Roseberry et al. (eds.), forthcoming).

The overall analytic framework that I shall be using can be seen in a developing tradition of genre analysis (Swales, 1990; Bhatia, 1993; Johns, 1997). I have drawn on these earlier frameworks for the categories and questions I shall use in this present study, perhaps giving greater emphasis to the distinction between *contextual* and *linguistic* analysis than has been the case in other studies. The two sections of my analytic framework contain a series of headings, which can, in turn, be expressed as a question or questions. The headings and their implicit questions are given in Table 7.2, and will be used in a detailed analysis of the RAT text.

It has been my experience that such an approach provides an useful basis for contextual and linguistic awareness raising during an EAP course, and offers a coherent basis for the development of curricula for writing instruction and the evaluation of written production.

CONTEXTUAL ANALYSIS

In the following section I shall use the questions which have been set under *Contextual Analysis* to give an account of the social/cultural dimensions of the example text.

1. Name
What is the name of the genre of which this text is an exemplar?

Without privileged knowledge, naming this short text is problematic. Six informants in Colombo, Sri Lanka (all English-medium-educated academics) were evenly split as to whether they would call it an *article* or a *report*, though, in the end, most felt more comfortable with *short report*. The fact that we know it is called a 'Commentary' helps us to situate the text as it can now be seen as standing in an analogous relation to similar sections in major journals such as *Nature* which also have space for short state-of-the-art reports of this nature. As Swales says: 'The genre names inherited and produced by discourse communities and imported by others constitute valuable ethnographic communication, but typically need further validation' (Swales, 1990: 58).

2. Social context
In what social setting is this kind of text typically produced? What constraints and obligations does this setting impose on writers and readers?

The article was written for publication in a specialist academic journal. In writing such a short piece, the author faces special constraints in terms of content and extent, but also has to meet normal academic standards of warrant and referencing.

3. Communicative purpose
What is the communicative purpose of this text?

Given that the piece was written at the invitation of the editors, the major explicit communicative purpose of this text must be to share recently established knowledge with a readership of peers. These special conditions minimise other subordinate purposes which may be associated with a published text – e.g. ensuring the professional standing of the author, or (possibly) challenging the reputation of another worker in the same field.

4. Roles
What roles may be required of writers and readers in this genre?

The readers and writers of this kind of newsletter have largely equal status as teachers or researchers, and the texts that are contributed to such journals are concerned with knowledge *forming* rather than with knowledge *transmission* (Myers, 1990, 1994). In such journals readers expect a high level of referencing

et under
ns of the

hort text
ri Lanka
re evenly
article or
comfort-
ıow it is
text as it
relation
ıs *Nature*
eports of
nherited
mported
:ommun-
(Swales,

specialist
iece, the
content
ıcademic

tation of
purpose
tablished
e special
ses which
ensuring
possibly)
er in the

ewsletter
earchers,
ırnals are
han with
In such
erencing

so that claims can be seen in their research context, and claims may often be tentative and tightly restricted. However, in the present instance, claims are made forcefully and unambiguously ('No biochemist had previously proposed . . .', '. . . enabled us to break the impasse reached with the EFE . . .', 'It is now clear that EFE is . . .', '. . . plant molecular biologists need not wait . . .'). This lack of tentativeness can be accounted for by the acknowledged significance of the breakthrough being reported, and by the status that Professor John and his colleagues' team has within the discipline.

While the role relationships instantiated in this text reduce its relevance for direct *modelling* by students (J. Flowerdew, 1993a), it makes the text interesting as a focus for analysis as it permits a useful discussion of difference between the present instance and more mainstream published articles.

5. Cultural values
What shared cultural values may be required of writers and readers in this genre?

Even though the text was not subject to the full rigour of peer review prior to publication, it nevertheless displays conformity to the full panoply of Western academic tradition. The author demonstrates his awareness of the imperative obligation to avoid plagiarism and to warrant all claims by·reference to empirical data or by citation.

6. Text context
What knowledge of other texts may be required of writers and readers in this genre?

Given the special nature of this particular text, the writer has more freedom of expression than he might have had if writing for another publication. Nevertheless, he demonstrates an awareness of the organisation and purpose of analogous texts in other journals (e.g. the short reports in *Nature* or *New Scientist*).

7. Formal text features
What shared knowledge of formal text features (conventions) is required to write effectively into this genre?

Apart from its use of normal paragraphing, the form of the conventional features of this text (citations, bibliography organisation, labelling or diagrams or figures) will depend on the 'house style' of the journal, and will usually be stated in explicit instructions to contributors.

CONTEXTUAL ANALYSIS: SUMMARY

This short text is a piece of written production from an established scholar in a specialised area of scientific research. The author gives an account of recent developments in his field to a peer readership. The specific form of his text is informed by its interrelations with other analogous short texts in major scientific publications. In the next part of our discussion we shall identify the extent to which the context of production has caused the writer to make lexico-grammatical choices which contrast with the choices that writers in other contexts would make – that is to say, the extent to which 'social context is predictive of text'? (Halliday, 1978: 189).

LINGUISTIC ANALYSIS

In an ideal world, the writing under consideration would be Part-Of-Speech (POS) tagged, thereby enabling a much more sophisticated analysis than is possible with raw text. However, there is also a value in seeing how far you can get with a bare minimum – so in this case I have worked with an ascii text file, a PC, *WordSmith Tools v.3.00* (Scott, 1996), and *MS Word 97.* These technical resources have been complemented by data on text differentiation in Biber, 1988 and Tribble, 1998 and a number of reference corpora (notably the two 1,000,000-word Written and Spoken data sets which will form the British National Corpus Sampler and a Romantic Fiction corpus derived from the Lancaster-Oslo-Bergen (LOB) corpus of British English).

Lexico-grammatical features

What lexico-grammatical features of the text are statistically prominent and stylistically salient?

Keywords

The first element in the analysis is to identify and investigate what Scott calls the 'keywords' in the text (Scott, 1997). Keyword analysis is carried out by:

(a) creating a wordlist with the *WordSmith Tools* Wordlist program.

(b) creating a keyword list by referencing this wordlist against a large corpus (in this instance a one-million-word list derived from the Written subset of the British National Corpus Sampler). Scott (1997) outlines the statistical procedures which underlie the program. The broad gist of Scott's explanation is that the Keywords program is able to sift out those words which are statistically prominent in the 'target' text (either outstandingly *frequent* or outstandingly *infrequent*) when compared with the frequencies of words in the reference corpus.

Keyword analysis offers one way of coming to grips with the choices our writer has made – in this first instance, the choices being in relation to

d scholar in
account of
cific form of
hort texts in
on we shall
d the writer
choices that
nt to which

t-Of-Speech
lysis than is
how far you
an ascii text
d 97. These
fferentiation
orpora (not-
ch will form
rpus derived
h).

minent and

at Scott calls
ried out by:

am.
large corpus
ritten subset
es the statist-
ist of Scott's
those words
er outstand-
with the fre-

choices our
relation to

Table 7.3 RAT keywords

N	WORD	FREQ.
1	EFE	12
2	MEMBRANE	7
3	ENZYME	7
4	ACTIVITY	11
5	ET	7
6	ETHYLENE	4
7	AL	7
8	MOLECULAR	4
9	SOLUBLE	4
10	HYDROXYLASE	3
11	FLAVANONE	3
12	YANG	3
13	ACC	3
14	BIOLOGISTS	3
15	HOFFMAN	3
16	PLANT	4

Swales's 'content schemata'. One of the specific claims that Scott (1997) makes for the Keywords program in *WordSmith Tools* is that it gives an insight into the 'aboutness' of a text. This claim is confirmed by the results in Table 7.3 where the top ten keywords of the RAT text are listed.

The list is a clear demonstration of the way in which membership of a particular disciplinary discourse community permits and requires the use of a range of content lexis which would be unallowable in genres catering for the needs of a broader readership. The occurrence of the apparently anomalous *et* and *al* is also accounted for by this membership. Indeed, I would predict that *et al* occurs more frequently in citations in science research than it does in the humanities where collective authorship is less the norm.

Not only do keyword lists give excellent insights into the 'aboutness' of single instances of a text – they can also be used in establishing a clearer understanding of the colligational and collocational relationships which generically significant words take on in the discourse. Thus a right sorted concordance for a single noun from the keyword list (see Table 8.4) tells us that *activity* can *depend, disappear,* or can *be required, required by,* or *associated with* various phenomena. It also tells us that it can be pre-modified by single and multiple nouns (*enzyme, flavanone 3-hydoxylase, lipoxygenase*).

Such information offers significant information for learners who wish to participate in the work of this discourse community. Keywords offer a rapid way of identifying the content lexis of a text, and also provide a means of gaining insights into broader text relations and stylistic choices.

Table 7.4 RAT 'activity'

```
    across the plasma membrane, and that activity depended on the maintenance
     belief in everyone's mind that enzyme activity depended on membrane integri
             ) of the in vivo activity, and their activity disappears completely when m
       because it was invariably found that activity disappears when tissue is ho
     Yang and Hoffman, 1984). Some EFE activity is retained by vacuoles isol
      sation of the flavanone 3-hydroxylase activity required inter alia anoxic c
     en went so far as to propose that EFE activity was associated with proton t
     cals which resulted from lipoxygenase activity (Yang and Hoffman, 1984). A
             itchell et al., 1988) of the in vivo activity, and their activity disappea
       t require Fe2+ and ascorbate for full activity, and it is now as amenable t
```

Frequency lists

While keyword analysis of a single instance of a genre provides an invaluable insight into the content schemata which inform a single instance of a genre,[2] it may offer fewer insights for other aspects of the lexico-grammar than one can obtain from a simple frequency wordlist. In order to extend the investigation of the grammar and style of this small text sample another procedure has been used.

(a) Create a frequency sorted wordlist for the text. Frequency sorted lists can be a useful starting point for text analysis (Tribble and Jones, 1997; Stubbs, forthcoming) as they give an immediate insight into words with prominent frequency.

(b) Generate left and right sorted concordances for each high frequency word (I have worked with the top ten in this instance). Left and right sorting of the contexts of the node word of a concordance is essential if patterning in language is to be identified (Tribble and Jones, 1997).

(c) On the basis of these lists and concordances, identify stylistically salient features of the text (Halliday, 1973; Leech and Short, 1981; Tribble and Jones, 1997; Tribble, 1998).

Initial results for the RAT text are already revealing. Although the most frequent words in a text (prominent) are not necessarily stylistically significant (salient), they provide a good way in to a text. Given in Tables 7.5 and 7.6 are the results for the 'top ten' words in the RAT text, referenced against raw frequency counts for the one-million-word spoken and written data sets used in the BNC Sampler, and a 24,000-word Romantic Fiction micro-corpus (Tribble, 1998) drawn from LOB.[3]

The first information we can get from these lists is that the text has the high percentage of definite nouns we associate with formal written discourse. This is evidenced by the contrasting percentages for definite article *the* for the four text sources I have sampled. Although percentages will not give information on statistical significance, they provide a useful rough means of differentiating between the texts in question.

Table 7.5 RAT frequency

RAT N	Word	Freq.	%
1	THE	35	5.96
2	AND	18	3.07
3	OF	18	3.07
4	IN	13	2.21
5	A	12	2.04
6	EFE	12	2.04
7	ACTIVITY	11	1.87
8	IS	11	1.87
9	TO	11	1.87
10	THAT	10	1.70

Table 7.6 Comparators

BNC Sampler: WRITTEN				BNC Sampler: SPOKEN				LOB Romantic Fiction			
N	Word	Freq.	%	N	Word	Freq.	%	N	Word	Freq.	%
1	THE	67,075	6.21	1	THE	38,962	3.71	1	THE	1,258	4.06
2	OF	32,656	3.02	2	I	33,478	3.19	2	TO	927	2.99
3	AND	28,900	2.68	3	YOU	27,334	2.60	3	AND	805	2.60
4	TO	26,680	2.47	4	IT	26,983	2.57	4	I	656	2.12
5	A	21,958	2.03	5	AND	26,013	2.48	5	A	633	2.04
6	IN	21,184	1.96	6	S	22,236	2.12	6	HE	575	1.86
7	IS	9,954	0.92	7	THAT	22,210	2.11	7	SHE	566	1.83
8	FOR	9,590	0.89	8	TO	22,142	2.11	8	HER	559	1.80
9	THAT	8,537	0.79	9	A	20,450	1.95	9	OF	533	1.72
10	WAS	8,362	0.77	10	OF	15,916	1.52	10	WAS	530	1.71

- RAT = 5.96% of all words
- BNC Written = 6.21%
- BNC Spoken = 3.72%
- Romantic Fiction = 4.06%

This impression is confirmed when the percentage of *of* is examined. Earlier studies (Biber, 1988; Biber and Finegan, 1989; Tribble, 1998) have demonstrated that a high frequency of *of* is frequently associated with formal written English, because as the frequency of *of* increases, so does its tendency to occur in the post-modifying structures typically found in nominally dense formal written English (Halliday, 1989). This is borne out in the present case by a concordance of *of* which demonstrates that the word occurs in post-modifying structures in 12 out of 16 instances in the RAT corpus text.

UK The final step in the **biosynthesis** of the plant growth regulator, ethy embrane systems) which were capable of converting ACC to ethylene, but . E, or at least a polypeptide component of the EFE. The amino acid sequencl s were also started up by the **discovery** of cell-free systems (always membra clear that EFE is a member of a group of soluble oxygenases that require onsible for catalysing the **hydroxylation** of 2S-flavanones to form 2R,3R-dihy t activity depended on the **maintenance** of a membrane potential (John, 1983 . It is now clear that EFE is a **member** of a group of soluble oxygenases th et al., 1991). What then is the moral of this tale? Quite simply that pl o enzymes, which took an enzyme out of a membrane where it was not loca lia anoxic conditions, and the presence of Fe^{2+} and ascorbate (Britsch and on fruits there was a complete **recovery** of the authentic EFE activity – as of the EFE. The amino acid sequence of this polypeptide resembled that se reached with the EFE. **Stabilisation** of the flavanone 3-hydroxylase acti nce of this polypeptide resembled that of flavanone 3-hydroxylase, a solub

(d) The concordance also revealed an interesting textualisation feature (marked with **bold text** in the concordance) which we will return to in the next section – this is the preference for nominalised structures over verbal structures when describing the processes involved in experimentation. Thus we find: **biosynthesis, discovery, hydroxylation, maintenance, recovery, stabilisation**. Again, this preference for nouns over verbs is typical of certain varieties of formal written English (Halliday, 1989).

Other features which arose from the raw frequency counts are consistent with the RAT text's formal written style – although the high frequency of *and* proved to be uninteresting as it is not the result of the employment of suasive rhetoric (one feature of which can be an unusually high occurrence of non-phrasal coordination[4]). Rather, it arises because of a choice about citation style (using the full form rather than *and*) and the fact we noted earlier that so many scientific papers are published under the names of more than one person. A further indication of the role of post-modifying prepositions as discriminators in written/spoken differentiation[5] comes in the data for 'in'. The percentage for *in* (2.21%) in the RAT text is comparable to the percentage in BNC Written (1.96%) – and contrasts with the percentages for BNC Spoken (1.29%) and LOB Romantic Fiction (1.27%), where, importantly, in most cases it follows a verb (or is part of a phrasal verb) or personal pronoun rather than post-modifying a noun, as shown in Tables 7.7 and 7.8.

The final comment we will make on the raw frequency data refers mainly to the use of verbs in the passage. Items 8 (*is*), 9 (*to*), and 10 (*that*) all offer additional insights into how verbs are used in the RAT text. Item 8 (is) occurs 11 times and has a higher percentage occurrence than BNC Written. We could predict that this is because the RAT text has a relatively high proportion of passive verb phrases[6] – and we would be correct. Of the 11 instances of *is*, five are agentless passives (see Table 7.9, where they are marked **P**).

embrane
at least a
o started
EFE is a
ysing the
d on the
EFE is a
he moral
out of a
e of Fe2+
f the au-
lypeptide
anone 3-
droxylase,

on feature
turn to in
tures over
perimenta-
intenance,
er verbs is
, 1989).

consistent
ncy of *and*
t of suasive
ce of non-
ut citation
earlier that
e than one
ositions as
ta for 'in'.
he percent-
es for BNC
ortantly, in
al pronoun
.8.
fers mainly
ut) all offer
n 8 (is) oc-
Written. We
gh propor-
1 instances
arked **P**).

Table 7.7 RAT 'in'

ng and Hoffman, 1984). All in all, little progress was bei
), which is readily assayed in vivo by supplying tissues wi
given rise to a firm belief in everyone's mind that enzyme
, pointing the biochemists in the right direction
erting ACC to ethylene, but in these preparations ethylene
orm 2R, 3R-dihydroflavonols in the biosynthetic pathway to
progress was being made in characterising EFE. para Unt
entered the scene. In 1990 Hamilton et al. reporte
identified a gene (pTOM13) in tomato which encoded for th
ied. para The final step in the biosynthesis of the plan
me had never been studied in vitro because it was invaria
chell et al., 1988) of the in vivo activity, and their ac

Table 7.8 Romantic Fiction 'in'

urtains and let the sea breeze in before he got into bed
o leave this house . . .' I broke in on her tirade. 'That's
a minute, Mrs. Landry,' he broke in gently. 'Loss of memor
act exactly the same,' she broke in. 'Please drive back. I
m the others is that my brother's in love with Lois. He nev
lking about when tea was brought in. Diana will soon be tw

Table 7.9 RAT 'is'

991). It is now clear that EFE is a member of a group of soluble oxyg
ant growth regulator, ethylene, is catalysed by the ethylene-forming e **p**
activity disappears when tissue is homogenised (Yang and Hoffman, 1 **p**
pletely when membrane integrity is lost (Porter et al., 1986; Mayne and **p**
onols and anthocyanidins. There is no obvious relationship enzymatically
Ververidis and John, 1991). It is now clear that EFE is a member of a
rbate for full activity, and it is now as amenable to biochemical stu
ene-forming enzyme (EFE), which is readily assayed in vivo by supplying **p**
ffman, 1984). Some EFE activity is retained by vacuoles isolated from le **p**
characterising EFE. Until, that is, the molecular biologists entered the
idis et al., 1991). What then is the moral of this tale? Quite simply

Additionally, of the 11 instances of *to*, five are associated with verb infinitives and two with passive structures (although one is accompanied by a non-animate agent):

biologists need not wait for their protein **to be characterised**
ethylene was shown **to be generated** by non-enzymatic reaction

That structures in the RAT text are also interesting – although again they lead us into questions of textualisation. Of the ten instances in the text, five are either introduced by a verb which comments on claims made by others:

Y *found* **that** . . .
X and Y *propose* **that** A, and **that** B . . .
X *reported* **that** . . .

or introduce a firm claim that the author is making:

It *is* clear **that** . . .

Lexico-grammatical features: conclusion

By examining a set of words that are statistically prominent in comparison with a general population of texts, along with a small number of high frequency words, it has been possible to identify the kinds of writing with which this text holds relations (formal written) and its content domain. We have also begun to see patterns in the text which contribute to the special identity of the text, and which almost certainly result from the constraints which the writer has had to respond to in order to ensure that the text is an allowable contribution to a specialist genre. The contextual analysis we have made connects with the lexico-grammatical analysis.

In the next section, we shall see to what extent other kinds of textual patterning can be identified, again using corpus linguistic tools for this purpose.

Textual patterning

Can textual patterns be identified in the text? What is the reason for such textual patterning?

We have already identified two kinds of textual patterning in the RAT text: the preference for nouns over verbs in describing processes, as was exemplified in the use of *biosynthesis, discovery, hydroxylation, maintenance, recovery,* and *stabilisation*; and the use of *that* clauses in the reporting of claims. A further example is offered here to demonstrate the potential of the kind of study we have been making in analysing a single text in relation to a large body of texts.

In this case, sentence beginnings were searched for (by looking, of course, for full-stops!). This time, however, an initial search on '*.' revealed a potentially interesting phenomenon at the *end* of sentences. Results are given in Table 7.10 which confirm this – this time based on a search for '*)'.

Out of 17 orthographic sentences in the RAT text, over 50% end with a parenthesised text citation – and another two such citations are found at coordinated clause boundaries. When compared with our reference texts, there is a striking contrast between Romantic Fiction, where (unsurprisingly) this phenomenon does not occur, and the BNC Written Sampler where it occurs 946 times (This feature is not relevant to the transcription conventions in BNC Spoken.) Space and time do not allow a detailed analysis of these results,[7] but a small sample indicates the stylistic and communicative parallels between this observed structure in the RAT text and its occurrences in the BNC.

Table 7.10 Warranting

```
1. a membrane potential (John, 1983). However this Mitchellia
2. activity (Yang and Hoffman, 1984). All in all, little progres
3. ogenised (Yang and Hoffman, 1984). Some EFE activity is r
4. ntegrity (Yang and Hoffman, 1984). The present author ev
5. n laboratory! (John et al., 1986). False trails were also
6. al, 1986; Mayne and Kende, 1986). para The earlier work
7. ate (Britsch and Grisebach, 1986). When these conditions
8. nzyme (Ververidis and John, 1991). It is now clear that EF
9. enzyme (Ververidis et al., 1991). para What then is the

1. f mesophyll (Guy and Kende, 1984) and by membrane vesic
2. iwifruits (Mitchell et al., 1988), but these systems ret

1. 8 or less (Mitchell et al., 1988) of the in vivo activity, a
```

Table 7.11 BNC final parenthesis

```
       been withdrawn' ( Knowles, 1978, 668 ). This reduction would have been
            al and rural-urban ( Robertson, 1961 ). This therefore leaves the u
     compensating counter-current ( Law 4 ). Until the recent repopulation
       ( what they did when they got there ). Until the mid-1970s these
         rising ( Johansen and Fuguitt, 1984 ). What evidence there is from
            eant ( Dean **9;426;hi et al., 1984a ). What was n't in doubt, th
     ve method ( Propst and Buyhoff, 1980 ). With regard to the search for
        eisure time ( Martin and Mason, 1976 ). Within this residue, which now
          onsiderably ( Glyn-Jones, 1979; 1982 ). Within the national parks'
      ferent groups of housing ( see fig 6 ). These are as follows: 1 ) CBD
     ere noted on a rough map ( see fig 3 ). This information was then
       the pumping processes ( in the dark ). C is thus negative for a finit
      edium ) is equal to zero ( modulo 2n ). Each such frequency is termed
```

What we can observe here is the way in which a final parenthetic element is used to:

- warrant claims or assertions by citing a published authority;
- clarify claims or assertions by reference to a figure or other part of the text;
- clarify claims or assertions with further comment.

In each of the instances expert writers are using a convention which does not occur at all in collections of student writing such as the LOCNESS[8] collection of student essays held at the University of Louvain (Granger, 1998a), or in a collection of Polish student essays from the Polish Corpus of Learners' English (PCLE).[9] As such it constitutes an aspect of the RAT text which may be genre specific, but is more likely to be common across a broad range of texts that are used in academic discourse communities.[10] It also indicates a starting point for student research. Asking learners to look at where and how parentheses are used is an excellent way of beginning an investigation of citation practices

– especially in that once the parenthesised citations have been identified, it is then easy to follow up how (and with which verbs, in which structures) the proper nouns which occur in such lists are used in the text.

Text structure

How is the text organised as a series of units of meaning? What is the reason for this organisation?

The final area to consider is the way in which information is organised as moves across the text. As Swales showed in the elaboration of his early CARS model for article introductions (Swales, 1981a, 1990), identifying such moves can be of great benefit to learners who are approaching a genre for the first time.

Although it operates at a more abstract level than CARS, the Situation–Problem–Response–Evaluation (SPRE) minimal discourse model proposed in Winter, 1977 and Hoey, 1983 also offers a powerfully generative way of viewing the relational patterns – and proves to be exceptionally apposite in analysing the four-paragraph text in question (refer to the Appendix for the full text). Thus the SPRE pattern maps closely on to the complete text, as is made clear by the opening sentence of each paragraph quoted below:

- **Situation**: outlines an earlier state-of-the-art in an aspect of microbiology – specifically the understanding of the role of a catalyst in biosynthesis and the research (*The final step in the biosynthesis of the plant growth regulator, ethylene, is catalysed by the ethylene-forming enzyme (EFE), which is readily assayed in vivo by supplying tissues with its substrate, 1-aminocyclopropane-1-carboxylic acid (ACC), but until recently the enzyme had never been studied in vitro because it was invariably found that activity disappears when tissue is homogenised.*)
- **Problem**: identifies the inadequate understanding on which this view was based (*The earlier work had given rise to a firm belief in everyone's mind that enzyme activity depended on membrane integrity.*)
- **Response**: describes a revised understanding in the light of research in another field (*In 1990 Hamilton et al. reported that they had identified a gene (pTOM13) in tomato which encoded for the EFE, or at least a polypeptide component of the EFE*)
- **Evaluation**: comments on the change in understanding and its impact on the field (*What then is the moral of this tale? Quite simply that plant molecular biologists need not wait for their protein to be characterised biochemically; molecular biology can be a very useful prelude to the biochemistry, pointing the biochemists in the right direction.*)

By combining a detailed analysis of the lexico-grammar of the text with this kind of account of the overall structure of the text, it begins to be possible to give a very precise (and well-contextualised) linguistic specification of an exemplar of a genre. As Stubbs says, 'the most powerful interpretation emerges if comparisons of texts across corpora are combined with the analysis of the organisation of individual texts' (Stubbs, 1996: 34).

RAT TEXT ANALYSIS: CONCLUSION

In the analysis so far we have seen how it is possible to use a genre analytic approach to identify the communicative context and purpose of a text, and then to develop a linguistic analysis (using corpus and discourse analysis tools) which makes it possible to identify the extent to which this context and purpose have shaped the linguistic choices the writer has made in realising the text. But how does this help us answer the questions with which this chapter started? What is the potential of a corpus approach to EAP writing instruction, and what corpus resources might be required in order to implement such an approach?

A corpus informed approach to EAP writing instruction?

In the introduction to this chapter I said that writers need four types of knowledge when producing allowable contributions to a particular genre: *content knowledge, writing process knowledge, context knowledge,* and *language system knowledge.* I hope that I have demonstrated in the sections above that genre analysis combined with corpus *tools* and corpus *resources* can make a contribution to the development of a writer's context knowledge and language system knowledge.

In the final section I would like to offer some further suggestions for developing what I shall call a corpus informed approach to EAP writing instruction.[11] I say corpus *informed* as it should by now be clear that a corpus (however well constructed) is not going to offer all the resources learners and teachers require. Content and writing process knowledge will remain areas for EAP teachers and students to address, however much contextual and linguistic investigation they might have done. What I am proposing, therefore, is the use of corpus resources in helping learners extend their understanding of written academic discourse – what Johns calls their *academic literacies* (Johns, 1997). The two areas on which I shall focus are (a) the resources which teachers might consider developing and (b) the ways in which they could use these resources to add a corpus informed dimension to their EAP programmes.

CORPUS RESOURCES,FOR EAP

Exemplar and analogue corpora

We have seen that one of the main tasks that EAP learners have is finding out about the kinds of genres they want to write into. They need to understand why texts are written in particular ways and what other texts they interrelate with, and they need to be able to use the linguistic resources which are associated with these genres. I have commented elsewhere (Tribble, 1998) that in the context of writing into a new genre 'difficult' can be interpreted as 'unfamiliar'. Most EFL writers in British universities are competent in the literacies required in their own academic cultures. Their problem is often that they do not know what the target performance looks like in English – often British academic genres are unfamiliar, and therefore difficult.[12] One reason for this is that it is often impossible (for example in the case of examination essays) to present learners with examples of the kinds of texts they are supposed to write as they are locked away in the university registry. Another reason is that the kinds of writing that undergraduates are asked to do *only* exist in educational institutions and have no clear, published analogue (Granger and Tribble, 1998; Tribble, 1997b).

One way, therefore, of developing corpus resources for EAP writing instruction is to collect texts which are the same as or, at least, analogous to the texts that your learners need to write, in other words, an *exemplar* corpus. Individual instances of such texts provide the basis for the kinds of genre analysis we have carried out here. Collections of such texts allow the learner to make generalisations about the genre which can further enhance their own written production (Bhatia, 1993). If it can be achieved, working with text collections built on production in the learners' target discourse community has many advantages, one of the greatest being that learners will be highly motivated to read the texts under consideration, and the generalisations which they are able to make on the basis of their analysis can be incorporated into their own written production. I have developed this kind of specialist corpus for project proposal writing (Tribble, 1998).

It is, however, remarkably difficult to assemble such collections, and it is often necessary to find *analogues* to the texts your learners wish to write, rather than specific exemplars (Tribble, 1997b; Johns, 1997). Tim Johns at Birmingham has taken this direction and built a corpus of *New Scientist* and science-related newspaper articles. Using this resource, he has developed a 'Virtual Data Driven Library' of activities and resources for self access use by students at Birmingham (http://sun1.bham.ac.uk./johnstf/ddl_lib.htm).

Exemplar corpora of specific genre production, and analogue corpora of close relatives of the target language production will constitute invaluable resources for language awareness raising and the investigation of grammar and lexis. Such corpora will be small and tightly focused. They will not, however, offer a basis for making generalisations about the language as a whole. Whatever the benefits of small exemplar and analogue corpora, it does not, however, mean that big corpora are bad things.

Reference corpora

Big corpora come back in to the picture when we consider the approaches to text and corpus analysis outlined in the discussion in the second part of this chapter. This approach does not depend on the compilation of a corpus of exact or analogous exemplars. Rather, it is based on the availability of a range of relevant reference corpora. The value of a reference corpus is that it permits the systematic comparison of individual instances of language use with language use in a general population of texts. This approach is inherent in much corpus linguistic work.

In developing programmes for EAP writing instruction which draw on reference corpora, teachers and learners will find themselves working with word lists and keyword lists developed by comparing examples of target language production with these large text collections. They will be identifying language use in one setting and comparing it with language use in other, clearly delineated genre contexts, and accounting for differences and similarities, and following lines of investigation similar to the one undertaken in this chapter. Working in this way, teachers will find that a combination of 'micro-corpora' of exemplar texts from their own disciplines and analogue corpora, studied in relation to large corpora such as the British National Corpus, will become increasingly important resources.[13]

CONCLUSION

In this chapter I have discussed the needs of student writers in EAP, and have attempted to outline how such students might be helped through a corpus informed approach to the analysis of written academic disource. Such an approach will draw on three kinds of corpora:

- *exemplar corpora* of texts directly related to the target writing behaviour of the learner;
- *analogue corpora* of texts which are similar to the target writing behaviour of the learner;
- *reference corpora* based on much larger text collections which will be used as a basis for the analysis of individual instances of genre production.

A corpus informed approach to writing instruction in EAP will make use of both big and small corpora, but will do so in a way which is sensitive to the needs of learners who are at different levels in their development of academic literacy. It will combine genre analysis with linguistic analysis in the early stages of EAP programmes when learners come to grips with the genres which will be important to them in their future academic careers. And it will continue to be relevant in the later stages of a programme as a means of extending the academic literacies that the learners require, and remediating problems which they discover as their confidence as writers increases.

NOTES

[1] My thanks to Ron White of CALS, Reading University, for permission to use this component of RAT.

[2] This is not the case if a larger sample of instances of a genre are studied. In this case, keywords can also give strong insights into the grammatical structure of a text (Tribble, 1998).

3. See Tribble, 1998 for a discussion of the potential value of these reference corpora.
4. A distinguishing feature of more suasive genres such as Project Proposals: see Tribble, 1998.
5. Biber's category 'Involved versus Informational Production' has four key components: written texts tend to have more *prepositions*, more *attributive adjectives*, a higher *type/token ratio*, and longer average *word length*.
6. Agentless passives are one of the major components of Biber's 'Abstract versus Non-abstract Information' dimension, which 'marks informational discourse that is abstract, technical and formal, versus other types of discourse' (Biber, 1988: 112–13).
7. But see Hyland (2000) for a useful discussion of citation in academic discourse.
8. My thanks to Professor Granger (granger@lige.ucl.ac.be) for permission to refer to the International Corpus of Learners' English (ICLE).
9. My thanks to Przemyslaw Kaszubski of Poznán University for permission to refer to PCLE.
10. See ch. 2 in Hyland (2000) for an extended discussion of citation in academic writing.
11. I shall assume that the reader will have access to basic technical resources – i.e. a modern PC, word-processing software (and basic skills) and a capable concordancing program. The most appropriate of these are *WordSmith Tools* (further details from Mike Scott: ms2928@liverpool.ac.uk) or *Monoconc Pro* (further details from Mike Barlow: barlow@athel.com).
12. As an extreme example, consider the 'problem' essay in legal studies.
13. It is relatively easy to use the header information to extract specific text collections from the BNC which can be saved separately from the main corpus and processed with a concordancing program.

APPENDIX 7.1: RAT CORPUS DATA (CONTACT PAUL THOMPSON [P.A.THOMPSON@READING.AC.UK] FOR FURTHER DETAILS)

How plant molecular biologists revealed a surprising relationship between two enzymes, which took an enzyme out of a membrane where it was not located, and put it into the soluble phase where it could be studied

The final step in the biosynthesis of the plant growth regulator, ethylene, is catalysed by the ethylene-forming enzyme (EFE), which is readily assayed in vivo by supplying tissues with its substrate, 1-aminocyclopropane-1-carboxylic acid (ACC), but until recently the enzyme had never been studied in vitro because it was invariably found that activity disappears when tissue is homogenised (Yang and Hoffman, 1984). Some EFE activity is retained by vacuoles isolated from leaf mesophyll (Guy and Kende, 1984) and by membrane vesicles in the juice squeezed from kiwifruits (Mitchell et al., 1988), but these systems retain only about one per cent (Porter et al., 1988) or less (Mitchell et al., 1988) of the in vivo activity, and their activity disappears completely when membrane integrity is lost (Porter et al., 1986; Mayne and Kende, 1986).

The earlier work had given rise to a firm belief in everyone's mind that enzyme activity depended on membrane integrity (Yang and Hoffman, 1984).

The present author even went so far as to propose that EFE activity was associated with proton translocation across the plasma membrane, and that activity depended on the maintenance of a membrane potential (John, 1983). However this Mitchellian EFE did not readily find experimental support, even from our own laboratory! (John et al., 1986). False trails were also started up by the discovery of cell-free systems (always membrane systems) which were capable of converting ACC to ethylene, but in these preparations ethylene was shown to be generated by non-enzymatic reactions between ACC and free-radicals which resulted from lipoxygenase activity (Yang and Hoffman, 1984). All in all, little progress was being made in characterising EFE. Until, that is, the molecular biologists entered the scene.

In 1990 Hamilton et al. reported that they had identified a gene (pTOM13) in tomato which encoded for the EFE, or at least a polypeptide component of the EFE. The amino acid sequence of this polypeptide resembled that of flavanone 3-hydroxylase, a soluble enzyme responsible for catalysing the hydroxylation of 2S-flavanones to form 2R,3R-dihydroflavonols in the biosynthetic pathway to flavonols and anthocyanidins. There is no obvious relationship enzymatically between flavanone 3-hydroxylase and the EFE. No biochemist had previously proposed an affinity between the two enzymes. Yet the structural relationship implied by the sequence homology provided the vital clue which enabled us to break the impasse reached with the EFE. Stabilisation of the flavanone 3-hydroxylase activity required inter alia anoxic conditions, and the presence of Fe2+ and ascorbate (Britsch and Grisebach, 1986). When these conditions were used to extract the EFE from melon fruits there was a complete recovery of the authentic EFE activity – as a soluble enzyme (Ververidis and John, 1991). It is now clear that EFE is a member of a group of soluble oxygenases that require Fe2+ and ascorbate for full activity, and it is now as amenable to biochemical studies as any other enzyme (Ververidis et al., 1991).

What then is the moral of this tale? Quite simply that plant molecular biologists need not wait for their protein to be characterised biochemically; molecular biology can be a very useful prelude to the biochemistry, pointing the biochemists in the right direction.

Prof. Philip John
Department of Agricultural Botany, Plant Science Laboratories, University of Reading, Reading RG6 2AS, UK