Science as a collocation. Phraseology in cancer research articles

Christopher Gledhill

GLEDHILL (1996) Science as a collocation. Phraseology in cancer research articles, in Botley S., Glass J, McEnery T. and A.Wilson (eds) *1996 Proceedings of Teaching and Language Corpora 1996* . UCREL Technical Papers Volume 9, pp:108-126.

Chris Gledhill.

Science as a collocation. Phraseology in cancer research articles.

Abstract.

English for Specific Purposes (ESP) is a specialist area of language teaching research that should benefit from corpus linguistics. Theory from corpus linguistics, such as the idiom principle (Sinclair 1987) can provide a powerful challenge to the intuitive areas of language use that have traditionally been of concern in the ESP syllabus (Sager et al. 1990). This paper marries a major theoretical backbone of ESP, the ethnographic approach of genre analysis (Swales 1990) to the large scale computational analysis of phraseology in a representative corpus.

The hypothesis explored in this paper is that written science is founded on a system of preferred expressions, and that collocation is a fundamental mechanism that allows for new formulations to take place throughout the text. On the basis of Johns and King's (1990) and Barlow's' (forthcoming) work on corpus-driven learning, I argue here that ESP is enhanced by an increased awareness of prototypical patterns and from the fact that deviation from the pattern is a key dynamic mechanism for the genre. Science is increasingly no longer seen as a body of facts transmitted via language, but as a special linguistic construct, mediated by the mechanisms of textual reformulation and phraseological innovation. Students of science should be aware of this process, and should be able to exploit the computational characterisation of normal expressions, that is: prototypes or 'preferred ways of saying things'. Putting it simply, new collocations are new scientific ideas, and knowing what constitutes a new collocation is a fundamental process in the acquisition of an ESP.

1 Introduction.

Stubbs (1996) has argued corpus analysis has not only provided linguists with a powerful tool for data, it has entailed a rapid rethinking of our basic assumptions about language. Stubbs' claim is that an underlying ideology corresponds with collocational patterns across large selections of authentic language:

> Representations are always from a point of view, and express group interests. Such points of view are not usually explicit, are often denied and may not be directly observable,

> because they are often a matter not of individual words , but of patterns of distribution and frequency. This is why we may need quantitative methods to study them. (Stubbs 1996: 235).

I would like to explore Stubbs' argument in an area of language that is well documented in linguistic research (Swales 1990, Myers 1990) but has received little attention from large scale corpus analysis: the language of research articles and in particular the collocational behaviour of grammatical items. I shall attempt to demonstrate that a thorough analysis of cancer research articles is possible using corpus analysis.

Sinclair's idiom principle essentially claims that lexis and syntax are co-selected features of language. The principle depends on the observation of patterns of language use across millions of words of authentic text. Recently Sinclair (1996) has emphasised the tendency in English to use fixed expressions with a very specific semantic, syntagmatic or pragmatic correlation. In particular, the pragmatic perspective has become more prominent, and has now been included as a feature of the second edition of the Cobuild dictionary (Channell 1993, Sinclair 1995). For example, using the predicative adjective *glad* requires a specific reason for having such a frame of mind. Compare: *I'm glad you're back* and *I'm happy you're back* (the reason is included in the statement). Such collocational scope has been termed 'prosody' in relation to semantic correlations (Louw 1993, Stubbs 1995) and I have used the term 'phraseology' in relation to the rhetorical, pragmatic force of a phrase (Gledhill 1995a). While Stubbs (1996) has claimed that semantic prosody reveals the ideological bias of single texts, it should also be possible to demonstrate this for a corpus representing the typical writing of scientists.

The idiom principle relies on probabilistic statements about the language. Idioms are merely typical expressions and may change over time. But new elements in an expression are likely to be interpreted in terms of the existing phraseology. I cite elsewhere (Gledhill 1995b) the example of *management* in cancer research. In a corpus of cancer research articles the term *management* only occurs in phrases such as '*patients received active management*'. Since this can only be

interpreted alongside more typical expressions such as '*patients received the pro-drug*, or *patients received drug X*' (where X is a treatment-related drug), we can assume that *management* is a technical term for a course of drugs. Elsewhere (Gledhill 1995b) I set out the ideological consequences of this formulation. In short, patients are never *given drugs* and are always expressed as active participants in the scientific process. But what is important to note here is that the expression contains a typical semantic prosody which extends to new or uncommon variations within that expression. This is the principle behind collocational frameworks (as in *a (quantity) of* (Renouf and Sinclair 1991)). And in the analysis of science texts, variation from the norm has been suggested as a way of introducing new metaphors from other scientific discourses (Pavel 1994) and as a way metaphors are expressed within the thematic development of expository text (Halliday and Martin 1993).

As Stubbs has pointed out, corpus analysis has brought with it a transformation in the way we see language. But since the majority of the work has been aimed at a 'representative sample' of the general language, corpus linguistics has only recently touched on specific varieties of English. Conversely, the rhetorically-oriented field of English for Specific Purposes (ESP) has not even begun to exploit corpus linguistics. Within the more established field of terminology, Thomas (1993) has conducted analysis of verbal complements in a medical English corpus and in discourse analysis Myers (1990, 1992), Kretzenbacher (1990), Salager-Meyer (1992) and others have analysed single grammatical or textual features of medical research articles (such as tense, passivity, lexical cohesion). But there have been no large scale discourse studies of lexical and collocational patterns in specialised corpora. Certainly, there has been no analysis of the collocational behaviour of grammatical items in these genres.

In this regard, Swales's (1990) notion of discourse community has been considered central to ESP. Swales rejects the traditional view of register, text defined by its own internal linguistic characteristics. Grammatical features for Swales are relative concepts, and variable in function depending on the genre. In turn, discourse communities are defined by their own discourse,

where they have their own internal mechanisms for self-regulation and communication, with their own pre-defined genres and tasks, their own lexis and jargon. Discourse communities adapt genres to their own purposes, and adapt linguistic resources such as rhetoric that is unique to the discourse community. But even within the scientific world, genres such as the research article exhibit a tremendous amount of variety. Genres evolve historically in terms of phylogenesis, as in Atkinson's (1988) work on the evolution of a research journal, textually in terms of logogenesis (Halliday and Martin 1993, Gledhill 1995a) and even within the linguistic competence of the individual, i.e. ontogenetically (Halliday and Martin 1993).

The kind of ethnographic approach that ESP represents can be applied to corpus linguistics. And is essential to take at least the rhetorical context into account when drawing up a specialised corpus. For example, a contextual analysis soon reveals that research articles are not the only genre used by the discourse community (grant proposals and monographs which have equal status) and that even within genre there will be variation in terms of the prestige of the journals and the relative rhetoric of experimental or theoretical articles. In a survey of fifteen cancer researchers (Gledhill 1995b), it emerged that a research article has a number of readings, depending on whether it is used indexically for facts or read in linear fashion, as a logical argumentation. The researchers also all had their own idiosyncratic views of cancer: cancer is a distributed, complex set of objectives and problems rather than a concrete concept. It is not one disease, but hundreds of related processes.

Clearly, ideology and the relative nature of the topic should be a central issue when interpreting our corpus data.

2 Collocations in cancer research articles.

> Knowing that your corpus is unbalanced is what counts. (Atkins et al. 1992:14)

The data in this article are based on a corpus of 150 research articles published in over 20

medical and pharmaceutical research journals between 1990 and 1993 (a breakdown is given in the appendix). The corpus is over 500 000 words long, a 'small' corpus in Stubbs' (1996) reckoning, but a corpus that is highly specific to one discourse community, one genre and one topic.

In order to provide balance in the corpus, I asked researchers from my survey to submit their own articles and also to recommend journals and even specific papers which they would consider relevant to their research (Gledhill 1995b). All the researchers had the ultimate goal of finding a cure of cancer, but their individual work ranged from finding drug cures (pharmacology and toxicology) to researching the properties of tumours (microbiology). Although all very specialised in content, the readership of journals ranged from the general one of the *British Medical Journal* (five articles), to one paper from the esoteric *Tetrahedron Letters.* Twenty articles came from the *International Journal of Cancer*. The resulting corpus is termed the Pharmaceutical Sciences Corpus (PSC) after the department at Aston university.

One purpose of the corpus was to analyse phraseology within different rhetorical sections. Thus the corpus was split into different sections:

TABLE 1: RHETORICAL SECTIONS IN THE PSC CORPUS.

Titles:	2 123 words (0.5% of the corpus)
Abstracts:	29 283 (6.6 %)
Introductions:	60 809 (13.7%)
Methods:	113 089 (25.5%)
Results:	123 084 (27.8%)
Discussion:	114 205 (25.8%)
TOTAL:	513 931 (100%)

Using Scott's (1993) Wordlist program, the content of the different sub-sections in the corpus can be compared to the corpus as a whole. Wordlist lists items that are significantly likely to occur in a specific sub-section, but not in the rest of the corpus. For example, *recently* occurs 52 times in the subcorpus of introductions and 102 times in the corpus as a whole (i.e. there are 50 instances elsewhere). Since introductions as a whole represent around 14% of the main corpus, a normal distribution would lead us to expect only 14 instances to occur, so this result is very highly significant. Items that are very significantly typical of specific sub-sections of the corpus are termed 'salient items' (Gledhill 1995b). The top twenty salient items in introductions are *Et, Al, Been, Has, Have, Introduction, Is, Recently, Studies, Cancer, Such, Genes, Effects, Variety, Can, Role, Report, It, We.* Some semantic tendencies are immediately apparent. Introductions are typical places for signalling recent research and other citations (*et al., studies, report, recently*) as well as introducing major topics (*cancer, genes*). Introductions also typically outline the empirical nature of research (*variety, effects, role*). But what is the role of the grammatical items, and why are so many of them salient?

Grammatical items are not distributed in a constant way across the corpus, and in Gledhill (1995b) it was grammatical items that were used for collocational analysis of different sections rather than lexical items. Such an analysis may be questioned as an intuitive pre-selection. But, as Moon (1987) demonstrates, grammatical items are often the most stable elements in longer idioms. The corpus work of Stubbs (1996) and others has centred on the analysis of lexical

keywords (such as *peace, work*). Even the second edition of the Cobuild dictionary has had to cut back on its original entries on grammatical items because they are so long. Instead, I argue that grammatical items are symptomatic of longer stretches of regular phraseology. It could also be argued that if a grammatical item stands before even low frequency lexical items, then there must be salient phrases that are prototypical for the subsection. For example, the analysis of grammatical items in introductions in the PSC corpus revealed the following expressions as prototypical phrases (the items analysed are underlined):

TABLE 2 TYPICAL PHRASEOLOGY OF SALIENT GRAMMATICAL ITEMS IN RESEARCH ARTICLE INTRODUCTIONS

p53 gene resistance has been reported (expression of report)
PIMO has received little attention (expression of report)
studies have shown that... (expression of report)
is an effective inhibitor (expression of evaluation)
(Compound X) is stable to the action of (Compound Y) (expression of chemical reaction)
alterations can be prepared (unmodified statement of fact)
use of agents such as dismutase (reformulating previous item)
it was also found that (reporting previous research)
In this study we examine (expression of report)
the purpose of the present study was to expand data (fixed and idiosyncratic expression)

This is a very limited selection. I have simply included the most frequent uses. But it indicates a preoccupation with expressing previous results and data (reporting verbs in the perfective) and signalling present preoccupations (use of *we*, projected clauses of purpose with *to*). Introduction sections also set out previously accepted facts (as shown with the use of *can* and the reformulating expression *such as*). In chemistry this involves an idiosyncratic expression: (adjective: *stable, unstable*) *to the action of* (compound involved in treatment).

Another finding from the corpus was that the same grammatical item would vary in usage depending on each section. Indeed, some grammatical items were salient across several sections, as the following table sets out (the first ten were selected from the Wordlist analysis):

TABLE 3: SALIENT GRAMMATICAL ITEMS IN THE PSC CORPUS

Titles: of, for, on, and, in (*no other grammatical items were salient*)
Abstracts: but, these, of, there, in, was, that, did, who, both
Introductions: been, has, have, is, such, can, it, we, of, to
Methods: were, was, then, at, for, each, and, from, after, with
Results: no, in, did, not, had, after, there, the, when, all
Discussions: that, be, may, is, our, in, not, this, we, have

While the specific phraseology of all these items is set out elsewhere (Gledhill 1995b), there has been no comparison of the phraseology of the same item across the corpus. This is an important point, since it may be that there are items that are either typical of a specific section (such as *such* in introductions) or typical of the corpus as a whole (such as *in*, in titles, abstracts, results and discussions). We can predict that *such* has a specific phraseology that is peculiar to introductions, but it is harder to say whether items such as *in* indicate a typical phraseology across the corpus or vary themselves depending on different sections. We shall therefore analyse *in* in some detail to determine its distributional behaviour. *In* is significant in that it is very typical of the corpus as a whole, as the following comparison with Cobuild shows (items that are significantly more frequent in the corpus are underlined: the Cobuild percentages are from Sinclair 1987):

TABLE 4: THE WORDLIST TOP TEN ITEMS IN THE PSC AND COBUILD CORPORA.

Rank	*Item*	*Tokens*	*PSC %*	*Cobuild %.*
1	the	29 122	5.8	6.1
2	of	21 309	4.3	3.0
3	and	14 610	2.9	2.8
4	in	14 349	2.8	1.8
5	a	8 631	1.7	2.4
6	to	8 125	1.7	2.7
7	was	6 146	1.2	1.0
8	with	3 543	1.1	0.6
9	for	5 224	1.0	0.8
10	were	5 162	1.0	0.4

To recap: *in* is salient in four rhetorical sections in the corpus and this presents us with the opportunity to use 'in' to test whether phraseology is truly variable in the corpus, or just at variance with the general language. In fact, we find below that its use does vary between certain sections. In addition, most of the PSC salient items are prepositions and auxiliary verbs (in contrast to items more frequent in Cobuild like 'that', as can be seen above), and this suggests that the research article genre differs from the general language at a basic grammatical level in areas such as prepositional and phrasal verb usage and in the construction and use of nominal groups.

In the collocational data below, I refer to four main semantic prosodies around which phraseological patterns appear to be organised: Clinical, empirical, biochemical and research-oriented phraseology. This classification is not fixed, but did emerge from the data so that in collocational frameworks the correspondence between a change in preposition often led to a different semantic prosody:

Clinical entities / processes: where the researchers carry out medical procedures (*cut, mix, separate, treatment, drug administration*)

Empirical processes: where the researchers evaluate or observe results (*increase, decrease, change, role, effect*)

Biochemical entities / processes: where the researchers identify some body part or process (*cancer, tumour, mice, gene expression, growth factor*)

Research processes: where researchers express their own discoursal activity (*findings, evaluation, prediction, suggest that, found that*)

2.1 Phraseology of *In* in Titles.

Here are the statistics which establish *in* as a very significant salient item in titles: 91 tokens (4.2% of the subcorpus) compared with 14 349 overall (2.9% of the whole corpus). The Chi square is 12.9 and this gives a probability of less than:0.0009. In titles 'in' has two specific roles:

1) as a prepositional phrase functioning as qualifier in complex nominals where the left collocate is a biochemical process. This is the most frequent use, and corresponds to the tendency for titles to be constructed as complex nominals (however 12% of titles include a predicator: i.e. a clausal element). While the left collocates of *in* is a biochemical entity or process, the actual heads of left collocate noun groups are usually an empirical or clinical items These are noted in bold below. Right collocates conform to the 'location' meaning of *in*, and tend to be related to disease:

changes in distribution of	cancer in	human liver
intake and risk of		children
improved detection of breast		women
determination of screening for		rats
surgical therapy of prostate		the elderly
gene	expression in	scrotal contents
gene		breast CYP1A1
receptor gene		cancer
gene		colorectal cancer
growth	factors in	gastric carcinoma
prognostic		breast cancer
Expression of trypsin and other		HB carcinoma
p53-like.., p53 **expression** and other		breast cancer
diethyl analogue	cell lines in	culture
growth-regulatory		a p53 pathway
human bladder cancer		protein

larger auxiliary	metastases in	obese women
colorectal adrenal		patients with (cancer)
breast cancer		melanoma
evaluation of...hepatic		patients
prediction of auxiliary lymph node		tumour-bearing animals

2) The second main use of *in* is as a postmodifying prepositional phrase where the left collocate is an empirical item whose statistical significance or medical potential is being emphasised: Again, right collocates tend to indicate disease but are clearly not locations:

Significant	change in	levels of specific in vitro residue
significant	changes in	cytokyne levels
highly significant		levels of stromal antigens
		cachexia mortality
		distribution of histogenic type

potential	role in	human disease
possible		the metastatic process
suggests a		tumor production
bioreducible drugs and their		cancer therapy

This second pattern is less prevalent in titles although there is an intermediate structure which includes a longer collocation involving the title salient item 'with'. The structure is: (modified empirical item X) in patients with (disease Y):

chemotherapy determination	in patients with	malignant melanoma
cell activation levels		
the function of folinic acid		terminal cancer
evaluation of pain measurement therapy		cancer of the liver
effectiveness of interferon alpha		
levels of coagulation factor		intraperitoneal malignancies

In summary, the first pattern for 'in' suggests a general semantic tendency for the qualifying phrase to specify the disease or the subjects in which the disease is to be found (mostly the 'spatial' meaning of *in* in traditional terms), while in the second pattern the completes the semantics of the left collocate.

2.2 Phraseology of *in* in Abstracts.

In occurs 912 times in the abstracts subcorpus (3.1% of the corpus) compared with 14 349 occurrences elsewhere (2.9% of the PSC). This gives a Chi square of 6.3 and a P score of 0.012 (still very significant). While in titles the spatial, biochemical use of *in* is most prevalent, in abstracts *in* is used most frequently in four patterns, almost all of them dealing with quantitative statements about data movement and results:

1) to modify nominal expressions of measurement (*significant increase in toxicity, reduction in levels, differences in cytotoxicity, decrease in uptake*).
2) as an particle in attributive or relational clauses (*accumulates in, is low in, resistance was narrower in the cell*), or as a phrasal element in research processes (*observed, detected*).
3) to post-qualify the expression of chemical or causal empirical processes (*role, resulted, used*).
4) introducing research with *this* (*in this study/ trial/ phase 1 study/ report...*).

In abstracts, *'in'* also introduces non-finite rankshifted clauses where given information on a chemical process is bundled in with the original information by explicative verbs such as *introduced, involved, implied* (as in: *this is a novel approach to adaptive resistance involved in*

the expression of ras oncogene). In contrast to its spatial meaning (*in the liver, in cells*) in titles, in the abstract this use is largely supplanted by a less specific meaning as in the use of *in + the +* (biochemical / clinical / empirical process), the most frequent of these involving the description of the mechanisms of carcinogenesis and tumour growth (*classification, suppression, treatment, transmission, dissemination, differentiation of the tumor, increase in the total number of cells*).

On the other hand, *in* is followed by zero-article in the case of 'problem' or 'disease-related' items: cancers, subjects or specific disease-related entities (*cancer, breast cancer, tumor-bearing animals, patients, tumor-bearing mice, cytokines, methylene chloride*). One explanation may be that just as article usage is highly idiomatic in certain specific semantic domains in the general language, it may be that phraseology becomes more polarised in the specific language.

In combines frequently in collocational frameworks (chains of collocation involving regular semantic classes) such as: *in the treatment of*. One particularly interesting premodifying term 'drug of choice' (6 occurrences) is also a frequent premodifier of *in the treatment of*' and this reveals the formation of a longer and relatively stable expression. involving the reformulation of similar concepts for new drugs:

(treatment X) is a (new) drug (commonly) used in the treatment of (disease Y):

aca C, a drug commonly used in the treatment of breast cancer patients
APD a commonly used drug in the treatment of cancer
(drug X) is a new H2 used in the treatment of cancer
(drug X) is a recent antagonist used in the treatment of gastric and duodenal cancer
(drug X) is a metallic antineoplastic agent that is used in the treatment of ... breast cancer
Harris et al. suggest the drug of potential value used in the treatment of ...tumours.

A collocational framework with *'of'* also introduces quantitative expressions in Introductions such as *in a variety of*, where the phraseology is a highly regular collocational framework. We find that the framework is involved in a longer phraseology: (biochemical process / entity or at times empirical process) is (used / empirical process) in (a) (wide) variety of- (treatment / disease related items):

Enzymes are involved	in a variety of	anticancer drugs
Both are inactivated	in a variety of	industrial drugs
Both are used as a solvent	in a variety of	industrial drugs
Splenic dl Plaz displays	a variety of	dysfunctions
the preclinical analysis	in a variety of	tumours...
antitumour efficacy	in a variety of	organs
Methyl chloride is used	in a variety of	consumer drugs
Methylene is used	in a variety of	pharmaceutical applications
macromolecules are used	in a variety of	formulations

2.3 Phraseology of *in* in Results sections

In results section, *in* is the most salient grammatical item of the subcorpus, with 3906 occurrences (3.3% of the subcorpus) out of 14349 (2.9% of the PSC corpus as a whole). The Chi square is high: 50.4, giving a probability of less than 0.0009.

'In' is used in three types of phrase in the results subcorpus. The most frequent expression indicates positive results which usually involve a higher score or increased amount in terms of measurement (*increase in, higher concentrations in*) and which usually indicate some comparison of figures. This is in contrast with abstracts, where *in* is seen to introduce expressions of data movement one way or the other (*decrease in, reduction in, difference in*). The second phraseology in results is closer to the essential spatial meaning of *'in'* in titles, indicating where a specific biochemical process was found / observed in the bodies of patients or subjects. Finally the third phraseology takes the form of a research process verb + preposition functioning as a cross reference to another section of the article (*as seen/ shown in*).

In the first 'positive results' pattern, the most typical uses of 'in' is with a statement of 'increase/s'

in data (61 occurrences) using either a biochemical process verb or a technical verb like 'yields, expressed, produced'. As with many relational processes in the corpus, the expression is most often modified by an evaluative epithet: (empirical process) (empirical evaluation) increase in (measurable, often disease-related empirical item):

treatment with butyrate	resulted in an increase in	relative tumor weights
2 weeks exposure	produced a linear increase in	the total number of.. tumors
exposure to methylene chl.	produced an increase in	incidence of renal dilation
treatment with... carcinogens	led to an overall increase in	alkaline phosphase activity
concentrations of deoxy..	expressed an increase in	the total tumor burden

Similar 'treatments' are involved in an expression which effectively becomes an idiom involving 'yielded' and a measurement item 'level'. Both of these items were seen to be frequent expressions in the abstract:

Treatment with dismutase	yielded modest increase in the levels of	lactase
butyrate-treated cells	yielded few increases in the level of	fetal matter
cells preexposed to butyrate	yielded an increase in the level of	spleen weight
treatment with cAMP	yielded a significant increase in the level of	...lesions
in vitro doses	yielded a similar increase in the levels of	...resorbsion

The second most frequent expression in the first pattern is the empirical process 'resulted in' where the direction of the data is emphasised by some intensifier and the observed phenomenon can also be a biochemical process: (clinical process) resulted in (intensifier) (empirical measure / biochemical process):

analysis	resulted in	marked	increases
protocols		significant	deaths
exposure to meth. chl.		70%	decrease
concentrations of dry MM		negative	induction
The same dose of DXR		strong	synergism
Since increasing the dietary BORA		total	loss of oral viability...

Another way of expressing positive results is to use a relational process verb with 'higher' where the phraseology is oriented around an evaluation of the change in data in animals or cells: (empirical relational process) (empirical measurement) higher in (animate material):

tended to be	higher in	dogs treated with 30mg
peak level is	markedly higher in	tumor cell lines
drug level is	consistently higher in	animals
leucocyte count is	significantly higher in	the liposomal DXR groups
5FU concentrations were	2 times higher in	animals necropsied at

This leads us to the second, spatial use of 'in' where the preposition introduces a biochemical entity. In some cases, as in the last examples, the biochemical entity is really a data set akin to the first use of 'in'. To give an example, 'in' can be seen in expressions of positive results where the data sets have derived from subjects or patients where there is comparison of 'in':

liver neoplasms were	more frequent than	in animals
drug levels were 30 times	higher than	in controls
significantly	higher levels than	in males
more typically	lower concentrations	in the corresponding control group
oxidised bases are present	at higher levels than	in those receiving liposomal drugs

A more typical spatial pattern involves technical biochemical processes including the classic clinical expression 'in vivo'. This use of *in* allows us to identify certain terminological limits which the phraseology must obey: a property of the language of cancer research which must presumably be acquired by those learning to write it. For example, certain biochemical processes, such as 'activity' usually only take place in 'organs':

cytotoxic	activity in	the organs
phosphatase		all the organs
PKC		cytosolic fractions
QK		various organs
antitumor		vivo

Similarly 'concentrations' are only found in 'tissues' or 'tumours':

variation of	concentration/s in	human tissues
relationship between 5FU		liver metastases
Data represent		murine tumors
x was the major metabolite		perfused rat liver
measurement of		tissues observed from the patient

The most frequent kind of materials to be found in biochemical entities are *proteins* (27 instances) which are typically found / examined in mammary cells:

examined the	protein/s in	normal mammary cells
found subcell location		mammary epithelial cells
the results show		epithelia; and fibroblast cells
detection of		tumor mammary cells
decreases the level of		breast tissue

Finally mutations are typically detected in genes (mutations in the p53 gene, in exon 6 of p53, in k-ras exons, in H-ras gene). This allows us to interpret k-ras exons as parts of the gene by analogy with typical expressions. An alternative wording is to premodify the mutation with a gene classifier, thus enabling it to be detected in tumours:

identification of ras mutations in	liver tumors
p53 mutations in	lung tumours
analysis of the p53 gene mutation in	methylene chloride-induces lung tumors
r-ras mutation in	case hepatomas
transcript mutation in	tumour-bearing animals

Interestingly, while we have noted that 'in vivo' is most often used as an adjunct (*studies were carried ou in vivo*), its complementary expression *'in vitro'* tends only to be used as a premodifier in noun groups, and so we get the following expressions:

The	in vitro	antitumour activity
The	in vitro	culture

useful in vitro growth
various doses of in vitro results
PKC activity of the in vitro system

Finally, the third pattern for *in* in results sections is the text referencing pattern, exemplified by the preposition's most frequent lexical left-collocate: 'shown in' (34 occurrences). The use of the present passive is noticeable in the following examples, since the past passive is generally reserved for empirical processes in Methods sections (*were increased, indicated, measured, determined,* ..):

Empirical measurement	Research process.	
results are	shown in	table X
results of the present study are		fig. X
correlations		
tumour response is		
the perfusate profiles		

A range of similar research-writing verbs fulfil a similar function:

clinical details are	detailed in	table X
samples are	given in	fig. X.
doses given are	illustrated in	
grain counts are	listed in	
these results are	plotted in	

The expression 'as shown by data in' also only refers to figures and tables. The only other expression where it is used in fact constitutes a very specific idiom which we observe in two structural chemistry texts, where the biochemical activity described in the methods section is referred to some result in a restricted expansion clause :

difference from controls	as seen in the first scoring event.
at this time point	
no change in esterase activity	
some intervals in rates	
significantly increased	

Conversely, the expression 'as described in' is uniquely used to cross reference to other sections of the research article, usually Methods, to indicate that the research process referred to is detailed there:

analysed for the presence of oxidised DNA bases as described in Methods
Incubation was carried out under conditions as described in Methods
tumours were examined histopathologically as described in the Methods
Q activity was determined as described in Materials and Methods
Accumulation was measured using... as described in Materials and Methods

The use of 'in' in conjunctive phrases is more varied than with other prepositions we observe in the corpus, and we note here briefly the expressions *in addition, in all, in comparison, in contrast*. These are compatible with the finding (Gledhill 1995b) that there is more explicit signalling in results sections.

2.4 Phraseology of *in* in Discussion sections.

In occurs 3991 times in the subcorpus (3.5%) compared with 14 349 occurrences in the PSC (2.9%). Chi square is 116.0, P score less than 0.0009 (very highly significant). To recap, in titles its left collocates were seen to be biochemical (metastases in, expression in, growth in) or empirical items (role of... in, change in). In abstracts, we noted a number of expressions involving empirical quantification (increase in, decrease in, reduction in, difference in). In results sections its use extended to positive quantification and comparison and cross reference to other parts of the research articles. In discussion sections the tendency is for empirical expressions of the shape of the data (the most frequent pattern, similar to its use in abstracts) and causal relations (the second most frequent pattern). A third pattern involves research processes, and a fourth comprises several expressions where 'in' is involved in a phrasal discourse marker. The latter three uses are unique to introductions and discussion sections.

Empirical items which denote general relationships or movement of data are the most frequent uses of *'in'* in discussions:

sensitive to the difference in peripheral substituents

there was no	proportions of t and o cells
This is likely due to the	charge distribution and geometry
This	cytotoxicity...
Results... complicated by global	biodistribution... of fragments

Other very frequent empirical data items (increase, change) are accompanied by empirical verbs such as 'resulted in', 'involved in', 'associated with' or research processes (such as 'was seen'). Another empirical item that signals causality forms an idiom: 'play a role in', where the presence of research or other empirical items is not obligatory, although some degree of evaluation and modality is often present (here emphasised in italics):

linkage does *not* play a major role in modulating the conformation of DNA
Our findings *suggest* that CsA might play an role in the differentiation of cells
Also, longbond structures *could* play an important role in other bond scission reactions
The phenopholyation of c143 TAA plays some role in the malignant proliferation of cells
accumulation of p53 alterations *may* play an important role in regulation of the cells

Similarly, biochemical items that are described as 'present in' others tend not to require expressions of empirical or research activity, and are stated as implicitly observed fact:

other transcription factors are present in these cells
other factors are present in the calf serum
p53 mutations were present in the majority of cancer cells
a small amount of contaminating mouse skin was present in the tissue
except for the 1464cm mode that is present in nearly all the resonance spectra

A similar pattern is seen in the expressions *is reflected in*, *is similar in*, and *is visible in*. The third pattern we note involves research processes, where a result is 'found' or 'observed', and this is similar to a pattern we noted in introductions sections (*similar response was observed in this study, LOH has already been found in all renal tumours*). The fourth pattern we note is a tendency for 'in' to be used in complex prepositions. These take the form of collocational frameworks where there is a similar discourse marker function throughout the pattern. For example, 'in... to' also allows for contrasts:

in	response	to	normal smooth muscle tissue
	addition		benign tumours
	contrast		benign smooth tissue and lymphomas

while 'in... with' signals that results have or have not been replicated elsewhere:

in	agreement	with	published data
	combination		other methylene results
	concurrence		Belleville et al.
	conjunction		the results obtained

The spatial use of 'in' as we have noted before reveals terminological consistency within the corpus. For example, only nude mice are used for skin grafts:

xenografting	in nude mice
in xenografts	
tumours xenografted	
inoculation or skingrafting	
The xenografts	

While frameworks with other common lexical items also reveal the collocational (and hence terminological) properties of tumors, cancer and carcinomas:

In....	benign tumour(s)	bladder cancer	colorectal carcinomas
	breast	breast	invasive
	clear-cell	colonic	
	colon	colorectal	
	colorectal	oesophageal	
	invasive	lung	
	malignant	pancreatic	
	p53-negative		
	primary		
	renal cell		
	Ta-Ti		
	various		

3 Conclusion

This paper has argued that a detailed and exhaustive analysis of grammatical items is possible in very specialised corpora. It has also attempted to demonstrate that even grammatical items as broad in usage and scope as *in* can reveal interesting patterns of discourse signalling, typical expression of terminology and the general relationship between broad semantic classes and syntactic variation.

As stated in the introduction to this article, English for Specific Purposes has been particularly interested in the analysis of research article sections, but has not provided a global analysis of typical expressions. Grammatical items were seen to vary considerably in usage between varying rhetorical sections of research articles. *In* was seen to play its traditional prepositional role as nominal postqualifier in titles, usually involving either biochemical or empirical relationships (*gene expression in, significant changes in*). In abstracts this role extended to expressions of quantitative statements about data movement and results (*differences in*) as well as facts about explaining the location of disease (*accumulates in, in patients*). In results this phraseology changed to the quantification of comparisons and positive results (*increases in, more stable than in*) and in post-verbal expressions of research (*as shown in*). Grammatically *in* moves from its

use in a qualifying prepositional phrase in titles and abstracts to prepositional phrase functioning as adjunct in results and discussions sections. Its role in collocational frameworks also changes throughout research articles, especially in results where the frameworks function as terminological groups, or expressions of research / empirical processes (*as described in, yielded a similar increase in*). In discussions these expressions are usually discourse signals (*in agreement with, in response to*).

We have attempted to demonstrate that variation of expression in a genre or a text type serves the pragmatic purposes of the individual producer or audience. Stubbs (1995) has consistently argued that changes in grammatical form are indicative of changes in ideology. While this article is not claiming that the ideology of science is transparently observable from the analysis of just one grammatical item, it could be claimed that this one analysis has touched on aspects that have traditionally been indicative of ideology. Thus, the mnor pattern for *in* in titles such as *in patients with* may represent a new tendency to express empirical evaluation and results in this section of the text. This would certainly be concomitant with current research that indicates the changing nature of titles (Jaime-Sisó 1993). As we saw above, this non-fixed view of phraseology is a basic premise of genre analysis (Swales 1990). Swales' (1990) point is that no grammatical feature has a stable function across a set of genres, and it is therefore unhelpful to define texts by their internal linguistic characteristics. Yet Swales and others have been unable to demonstrate these patterns, largely because of their preoccupation with rhetorical macrostructure. Swales' own *ad hoc* analysis of what are typical grammatical features in research article introductions is thus hampered by this point: as Stubbs is eager to point out (Stubbs 1996). Instead, the analysis presented above would confirm Swales' point more convincingly on genre. It also demonstrates the inadequacy of register-based analyses such as Biber (1988) and Krezenbacher (1990), which tend to associate similar functions to signal linguistic features.

Generally, the use of *in* reveals not only the typical expressions but also the rhetorical preoccupations of each section. I argue that these expressions are fundamental and unique to this

genre. But at least from my own survey, researchers are generally unaware of such patterns: certainly no editorial policy could explain them all and none of the researchers (some of whom were non-native speakers) were aware of even basic patterns (such as *patitents receive drugs*). In addition, the consistent correspondence between collocational frameworks and distinct semantic categories (such as biochemical, research-oriented etc.) are fundamental organising principles for the phraseology of science writing. This is indicative of a body of knowledge, a discourse which has to be progressively learnt by the novice. The kind of corpus analysis we set out above may be used as a way of revealing terminological regularities to which students and researchers would not normally have access. For example, the word *cancer* itself is always premodified by the location of the cancer to indicate its type, while *tumour* is usually premodified by an expression of its quality (*benign, invasive*). At the very least, medical students (that is learners of English for Specific Purposes) and syllabus-designers would benefit from an awareness of the prevailing regularity of these expressions.

References

ATKINS S., CLEAR J. and OSTLER N. 1992 "Corpus design criteria." in Literary and Linguistic Computing Vol. 7/1 :1-15

ATKINSON D. 1992 "The evolution of medical research and writing from 1735 to 1985: the case of the *Edinburgh Medical Journal*" in Applied Linguistics Vol. 13/4: 337-374

BARLOW M. (forthcoming) 'Corpora for theory and practice' submitted to Journal for Literary and Linguistic Computing.

BIBER D. 1993 "The multidimensional approach to linguistic analyses of genre variation: an overview of methodology and findings." in Computers and the Humanities Vol. 26 :331-345

CHANNEL J.C. 1993 The coding and extraction of pragmatic information in a dictionary data-base. Unpublished, Cobuild.

FRANCIS G. 1993 "A corpus-driven approach to grammar." in Baker et al. (eds.) 1993 :137-156

GLEDHILL C. 1995a "Collocation and genre analysis. The discourse function of collocation in cancer research abstracts and articles." In Zeitschrift für Anglistik und Amerikanistik. Vol. 1/1995:1-26

GLEDHILL C. 1995b "Scientific innovation and the phraseology of rhetoric. Posture, reformulation and collocation in cancer research articles". Unpublished PhD thesis, University of Aston.

HALLIDAY M.A.K. 1985 Introduction to Functional Grammar London: Edward Arnold

HALLIDAY M.A.K. and JAMES Z.L. 1993 "A quantitative study of polarity and primary tense in the English finite clause." in J. McH. Sinclair (et al.) 1993 :32-66

HALLIDAY M.A.K. and MARTIN J. 1993 Writing Science: Literacy and Discursive Power London: Falmer Press

JAIME-SISÓ M. 1993 "The new role of titles in research articles." unpublished paper

presented at the 5th International Systemic Workshop on corpus-based studies, Universidad complutense de Madrid, 26-29 July 1993

JOHNS T. and KING P. 1993 Data-Driven Learning Workshop presented at the Baleap meeting, University of Birmingham, March 22 1993

KAPLAN R. and GRABE W. 1992 "The fiction in science writing." in Schröder (ed.) :199-217

KRETZENBACHER H.L. 1990 Rekapitulation: Textstrategien der Zusammenfassung von Wissenschaftlichen Fachtexten Tübingen: Gunter Narr Verlag

LOUW B. 1993 "Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies." in Baker et al. (eds.) 1993 :157-176

MARTIN J.R. 1991 "Nominalization in science and humanities: Distilling knowledge and scaffolding text." in E. Ventola (ed.) 1991 :307-337

MOON R. 1992 "There is reason in the roasting of eggs. A comparison of fixed expressions in native speaker dictionaries." in Euralex '92 Proceedings Oxford University Press :493-502

MYERS G. 1990 Writing Biology: Texts in the Social Construction of Scientific Knowledge Milwaukee: University of Wisconsin Press

MYERS G. 1991 “Lexical cohesion and specialized knowledge in science and popular science texts.” in Discourse Processes Vol. 14/1 :1-26

PAVEL S. 1993a "Neology and phraseology as terminology-in-the-making." in H.B. Sonneveld & K.L.Loening (eds.) 1993: 21-34

PAWLEY A. and SYDER F.H. 1985 "Two puzzles for linguistic theory: naturelike selection and naturelike fluency." in Richards and Schmidt (eds.) 1985 Language and Communication London: Longman

PETERS A. 1983 The Units of Language Acquisition Csambridge: Cambridge University Press

RENOUF A. and SINCLAIR J. McH. 1991 "Collocational frameworks in English." in K. Aijmer and B. Altenberg 1991 :128-144

SAGER J.C. DUNGWORTH D. and P.F. McDONALD 1980 English Special Languages: Principles and Practice in Science and Technology Wiesbaden, Oscar Nadstetter Verlag

SALAGER-MEYER F. 1990b "Discoursal Flaws In Medical English Abstracts" in Text Vol. 10/4: 365-384

SCOTT M. 1993 ""Lexical tools for genre analysis for computers." Unpublished MS presented at the BAAL annual meeting 14-16 Sept. 1993

SINCLAIR J. McH. (ed.) 1987a Looking Up: An Account of the Collins COBUILD Project London: Collins ELT

SINCLAIR J. McH. 1991 Corpus, Concordance, Collocation Oxford, Oxford University Press

SINCLAIR J. McH. 1996 "The chains of choice." Unpublished paper presented at the 'The data of linguistics' Workshop, University of Salford 16 March 1996. North-West Centre for Romance Linguistics.

SONNEVELD H.B. and LOENING K.L. (eds.) 1993 Terminology. Applications in interdisciplinary communication. John Benjamins: Amsterdam.

STUBBS M. 1994 "Grammar, text and ideology: computer-assisted methods in the linguistics of representation". in Applied Linguistics Vol.15/2 :201-223

STUBBS M. 1996 *Text and Corpus Linguistics*. Oxford: Blackwell

SWALES J.1990 Genre Analysis: English in Academic and Research Settings Cambridge: Cambridge University Press

WEINGART P. 1993 "Science abused? Challenging a legend". in *Science in Context* Vol. 6/2 : 555-567

WIDDOWSON H.G. 1989 "Knowledge of language and ability for use." in Applied Linguistics Vol 10/2

WILLIS D. 1990 The Lexical Syllabus London: Collins ELT