The Discourse function of collocation in research article introductions

Christopher Gledhill

PERGAMON English for Specific Purposes 19 (2000) 115–135

ENGLISH FOR SPECIFIC PURPOSES

www.elsevier.com/locate/esp

The discourse function of collocation in research article introductions

Chris Gledhill

School of Modern Languages, Buchanan Building, Union St., St Andrews, Fife KY16 9PH, Scotland, UK

Abstract

The increasing use of computer-held text corpora containing many millions of words has allowed linguists to establish lexico-grammatical patterns in language that were previously unavailable to observers. Such patterns range from lexical collocations and idioms to the phraseology of grammatical items. Recently, collocations of high frequency words in medical research abstracts and articles have been found to be useful indicators of the prototypical phraseology of the genre. In this article we characterize the phraseology of Introductions from a corpus of 150 cancer research articles. We explain the fixedness and idiosyncratic nature of scientific phraseology in terms of discourse processes such as reformulation. We argue for the design of a representative and specialized corpus of the research article and a contextual approach to corpus work that is appropriate to the teaching of languages for specific purposes (LSP) and the ethnographic aims of genre analysis in general.

Keywords: Discourse; Genre; Computer analysis; Phraseology

1. Introductions

This article proposes a computer corpus-based methodology to describe the phraseology of the research article genre. The assumption is that language is organized in terms of a lexico-grammar (Halliday, 1985; Sinclair, 1991). This perspective emphasizes the idiomatic nature of language, especially the dependent relationship between the vocabulary and the grammatical system. Vocabulary

PII: S0889-4906(98)00015-5

items are not always single items or simply "content words". They can involve multi-word units, such as idioms, clichés or fixed expressions which have both a consistency of form and of meaning (Cruse, 1984). There have been few attempts to characterize idioms and fixed expressions in scientific texts, and the role that such items might play in these texts has not until recently been an issue of much debate. Although there are few traditional idioms as such in a specialized corpus, it can be seen that much of the language involved in research article Introductions is idiomatic and highly stereotypical in nature. Since the purpose of computer-based corpus analysis is to look for patterns which would not ordinarily be the focus of genre analysis, we focus particularly here on the typical contexts (collocations) of grammatical words. We attempt to account for their role in terms of the textual function of collocation and the role of fixed expression in the discourse community.

2. Corpus and genre analysis

In recent years corpus linguistics has enjoyed an explosion of interest (Barnbrook, 1996; McEnery & Wilson, 1996; Stubbs, 1996; and others) thanks to the widespread availability of computer-held data-bases of texts. Among many applications, corpus analyses have attempted to describe dialects and registers of English from sizable text collections (Biber, 1986; Altenberg & Eeg-Olofsson, 1990; Aijmer & Altenberg, 1991; Biber & Finegan, 1994). However, as McEnery and Wilson (1996) note, most corpus work on English to date has concentrated on very broad sweeps of language. Although in applied areas such as terminology and lexicology there has been considerable exploration of phraseology in technical corpora (Sager, Dungworth & McDonald, 1980; Thomas, 1993; Pavel, 1993), work in English for Specific Purposes has only recently taken the opportunity to explore large corpora, mostly in terms of rhetorical structure (Thetela, 1997; Bittencourt dos Santos, 1996) but also exploiting a lexico-grammatical perspective (Banks, 1994a,b; Gledhill, 1996, 1997). The attraction of a combined approach to both genre and corpus analysis lies in the potential for a corpus to reveal recurrent patterns across a representative sample of texts. The genre approach in turn allows us to nuance the often monolithic descriptions that may emerge from corpus work, by offering a contextual, ethnographic basis for the construction of a textual corpus as well as a view of text as a series of choices, ebbing from one style to the next.

Many ESP studies have been carried out with single grammatical categories such as hedging, negation, passivization and so on (Biber & Finegan, 1994). To our knowledge, however, there has been no analysis of the distribution and collocational behavior of idioms and lexical items in different sections of the academic research article. Terminologists have perhaps the most established view of collocation in science, but usually this has concentrated on the neological construction of complex nominals or verb-complement relations (Thomas, 1993;

Pavel, 1993). In this article we specifically target grammatical words rather than lexical items or grammatical categories. We argue below that the analysis of grammatical words is an efficient way of arriving at a description of the most typical expressions in the corpus.

One insight of the lexico-grammar perspective (Sinclair, 1991) has been to question the traditional idea of the word class and to argue instead that there is a cline, with high frequency polyvalent items at one end (''grammatical words'' are polyvalent because of the wider variety of words they collocate with) and highly specialized items at the other (''lexical words'' whose collocational context is more likely to be restricted). In between both ends of the spectrum lexicologists recognize lexical complexes with varying degrees of semantic and syntagmatic cohesion: idioms (''kicked the bucket''), polywords (''by and large''), sentence frames or ''prefabs'' (''the fact that'') and collocations. Collocations can be bound as in ''foot the bill'', ''shrug one's shoulder'' (where one lexical word is obligatory), or unbound as in ''a [time expression] ago'' or ''[something negative]'' sets in. It has been noted that grammatical words play a particularly important role in the cohesion of these expressions, even where lexical items are seen to be central. This is particularly true of collocational frameworks of the type ''a [quantity] of'', ''too [relative time] in the [time expression]'' (*a bucket of, too late in the day*, etc.) (Renouf & Sinclair, 1991). We set out to discover the role of these intermediary elements in scientific texts. Not surprisingly, there have been few analyses of the phraseology of grammatical words because even a small corpus produces a seemingly unmanageable amount of data. But there is also perhaps a lingering belief that grammatical words reveal less interesting data. Halliday and Hasan (1976) once appeared to rule out a role for high frequency items in textual cohesion. But there is now copious evidence to suggest that high frequency items have a restricted and idiosyncratic syntax (Sinclair, 1991) and that they are embedded in set phrases which have rhetorical force (Moon, 1992). Moon points out that since highly marked expressions exist in conjunction with less marked, less idiomatic wordings (compare ''the negotiations progressed at a snail's pace'' with ''the negotiations progressed slowly''), they are used in situations where the speaker intends a new level of evaluation. While these observations were based on the general language, it is clear that an analysis of scientific discourse would benefit from an approach which could systematically determine the collocational properties of even high frequency words. It would also be useful to review the role of fixed expressions within science writing itself. For example, our corpus indicates that the expression *in vitro* is mostly involved as a single unit functioning as classifier (*in vitro* determines a type of experimental procedure as in *in vitro* fertilization) as opposed to *in vivo*, which tends to occur more as an end-of-sentence adjunct (as in *X* were performed *in vivo*). Such tendencies presumably reflect the working practice of empirical science.

The main hypothesis in this article is that by exploring broad grammatical features of scientific discourse, we can trawl for these frequently expressed formulations. This would establish a typical or generic phraseology for an overall description of the genre. As we have mentioned, there is a dearth of information

on the textual properties of collocation. When phraseology is analysed at a textual level, it is often identified with rhetorical purpose. For example, in his analysis of oceanography texts, Banks, 1994b correlated the distribution of the passive, personal pronouns and modality in verbs and adverbs across rhetorical sections. Interestingly, he notes that the lexical hedging of verbs by modals (*can, may*) is so widespread towards the latter part of research articles that their effect may be redundant. This suggests that a conventional "voice" has become entrenched in some science writing, and Myers (1989) has argued that such obligatory expressions are an imposition of the discourse community (including preferred expressions for *claims, denials, coining of new terms, apologizing for speculation*). Our own survey of science writers (Gledhill forthcoming) suggests that they are largely unaware of these specific phraseological conventions, despite evidence that the conventions are well established.

3. The pharmaceutical sciences corpus (PSC)

It is now widely accepted in corpus linguistics that the context of a specialized corpus must be as explicit as possible and must display clear design criteria. The corpus analysed here was designed on the basis of an initial ethnographic survey of 15 researchers at Aston University, U.K. They were all involved in cancer research, although they had different objectives and declared themselves from several disciplines (from microbiology to structural chemistry). This kind of group is a very loose discourse community, linked by institutional ties of teaching and common overall research patterns. The overall goal of a cure for cancer was far from their immediate objective, and indeed their definitions of cancer as a concept were extremely varied. What is intended to emerge from our corpus therefore is the phraseology not of a small group of researchers but of a research paradigm that spans at least 22 journals and a wide but interrelated set of specialisms.

In Gledhill (1995a; and forthcoming) we set out the details of a survey of the Pharmaceutical Sciences department, as well as the criteria for selecting 150 research articles for the corpus (over half a million words). In summary, the texts were selected according to the following criteria:

1. Accessibility (some texts were derived electronically from electronic research databases such as ADONIS and specific texts were selected on the basis of the researchers declared areas of interest).
2. Prestige (research articles were requested from journals such as *Journal of the National Cancer Institute* because more than one researcher said they saw these as key texts and also from the highest ranking cancer/oncology journals in the 1989 Science Citation Index such as *Cancer Chemotherapy and Pharmacology*).
3. Authorship (fifteen texts were made immediately available by the researchers themselves, thus making the corpus slightly more representative of production as well as reception. This gave a range from the more popular *Trends in Pharmaceutical Sciences* to the esoteric *Tetrahedron Letters*).

Table 1
Constitution of the pharmaceutical sciences corpus (by words and subsections)

Subcorpus	Code	% of PSC	Tokens
Title	T	0.5%	2127
Abstract	A	5.8%	29,136
Introduction	I	11.8%	59,724
Methods	M	27.5%	137,161
Results	R	27.6%	119,746
Discussion	D	26.7%	114,829
Total		100%	499,370

When scanned with the publisher's permission, the whole text of an article was included in the corpus, although references and the not inconsiderable amount of text which accompanied diagrams were excluded. Half of the corpus represents cancer-oriented journals (74 research articles) with another half from medicinal chemistry (76 research articles). The content of the corpus can be gauged from Appendix 1.

Once collected, the corpus was split into sections: Title—Abstract—Introduction—Methods—Results—Discussion. Although the proportion of sections varied according to the journal and article, Table 1 indicates the average size of each (for simplicity, hybrid sections have been removed from these figures).

The Keyword computer program now incorporated in the Wordsmith program, available at web site http://www.oup.co.uk/elt/software/wsmith) was used to compare frequency lists from the corpus, providing a list of frequent words (salient items) that were more significantly frequent in one section than in the rest of the corpus. This enabled a principled approach to deciding which grammatical words to analyse. Salient items are therefore an internal measure, typical of the rhetorical section rather than of the corpus as a whole. The salient grammatical items for the six main rhetorical sections in the corpus are listed in the table below (statistics for each section are provided in Gledhill forthcoming). It should be noted that only five grammatical items are statistically significant in titles as compared with the corpus as a whole (Table 2).

Salient items in Introductions (with the data that motivate their selection and phraseological summaries) are analysed in detail in the next section. The next stage involved contextual analysis of each grammatical word using Microconcord where the words were aligned for ease of analysis and the most frequent collocates (words occurring left and right of the main word) were calculated. In the results below, we have limited the number of examples of collocation to five.

Since there is no established metalanguage or system of notation for collocational analysis, we use the following conventions. The word form under analysis is underlined in the text. Italics are reserved for corpus citations not used in concordances. We use triangle brackets for fixed collocational units (such as ⟨have been⟩ ⟨was to⟩) and underline obligatory collocations that are several words away (⟨it is⟩ (important) to). We use square brackets as a short hand for free

Table 2
Salient grammatical words in the pharmaceutical sciences corpus

Rank	Title	Abs.	Intro.	Methods	Results	Disc.
1	of	but	been	were	no	that
2	for	these	has	was	in	be
3	on	of	have	at	did	may
4	and	there	is	then	not	is
5	in	in	such	for	had	our
6	—	was	can	each	after	in
7	—	that	it	and	there	not
8	—	did	we	from	the	this
9	—	who	of	after	when	we
10	—	both	to	with	all	have

collocations which display a consistent semantic content or prosody, the four most common in the corpus being biochemical, clinical, empirical, research-oriented. Research article sections are given capitals, i.e. Abstracts, Introductions, and so on.

4. Phraseology in PSC introductions sections

The PSC Introductions subcorpus contains 59,724 words (just over 10% of the total corpus). The Wordlist comparison with the whole corpus gives the following data (only the first ten salient grammatical words were selected) (Table 3.

While all the words prove interesting from a phraseological perspective, we narrow our analysis below down to the verbs (has, have, been, is) and prepositions (of, to). The other items have very specific phraseologies which we can summarize here. Such plays an important role in Introductions by reformulating biochemical processes as hyponyms within a set taxonomy (*antitumour agents such as NMU, use of hormonal enzymes such as dismutase,...such an inhibitory agent*). The auxiliary can serves not to modalize or signal hedging but to express potential clinical processes (*methods can be considered, alterations can be prepared*) or to explain a biochemical's "ability to" operate in a novel way involving a technical or sub-technical verb (*gene products can dimerize, cytokines can flip*). Although can is an important item for explaining innovations, it is not salient in Discussions and is seemingly replaced by *X may be shown to Y*. We is used in the expression of conviction in Discussions (*we conclude that*), whereas in Introductions reference to the authors is primarily to express the rhetorical move "occupy the research gap" (Swales, 1990) (*We have in this article studied NAK cell susceptibility, we have in this report studied tumor-drug distribution*). Unsurprisingly, it is used in empty extraposed clauses throughout the corpus (its pronominal use is minimal overall) and the word is clearly very salient in Introductions as part of the phraseology of have, been and to as we see below.

Table 3
Salient grammatical words in research article introductions

Word Rank (Intro.)	Word.	Frequency (Intro.)	Proportion (Intro.)	Frequency Main corpus	Proportion Main corpus	χ^2 score	Probability[a]
3	BEEN	346	(0.6%)	966	(0.2%)	341.1	0.000
4	HAS	283	(0.5%)	741	(0.1%)	310.3	0.000
5	HAVE	359	(0.6%)	1127	(0.2%)	285.4	0.000
7	IS	643	(1.1%)	3169	(0.6%)	156.3	0.000
11	SUCH	113	(0.2%)	388	(0.001%)	73.7	0.000
15	CAN	120	(0.2%)	468	(0.001%)	58.1	0.000
18	IT	207	(0.3%)	1006	(0.2%)	52.2	0.000
19	WE	200	(0.3%)	972	(0.2%)	50.4	0.000
25	OF	2874	(4.8%)	21,309	(4.3%)	41.4	0.000
32	TO	1233	(2.1%)	8631	(1.7%)	36.6	0.000

[a]A probability of 0.000 indicates very high statistical significance.

4.1. Has/Have/Been

Has been/have been are used in two types of perfect passive construction which have been identified as typical of reporting in Introductions (Salager-Meyer, 1992). In cancer research texts the phraseology is almost exclusively involved with establishing a connection between a drug or biochemical process and a disease. There is a very distinct correspondence between clause type and semantics: to-clauses (these are "projecting" clauses, where the initial clause labels the next clause as a statement, idea or fact e.g. "*the drug has been shown to be*...") and extraposed that-clauses (extraposed clauses project a research idea or fact through an empty subject as in "*it has been thought that*"...). Of the two structures, that-clauses are more frequent in Abstracts and Discussions (as can be gathered from Table 3, above) where the function of reporting present findings is more evident (the most frequent being: *we have demonstrated that, these findings indicate that*). From the examples below, it appears that the to-clauses emphasize the agent of some biochemical action, while the that-clauses are oriented around ideas. In Introductions, the more frequent to-clause pattern establishes the research reporting of the various biochemicals in the article using projected verb-complement clauses [biochemical entity] ⟨has been shown to⟩ (32 examples in Introductions) for example, *TNF alpha has been shown to deliver the toxicity of ricin A*. The second most frequent pattern is biochemical process, usually involving a treatment ⟨has/have been reported to⟩ plus a descriptive projecting clause, for example, *CsA therapy has been reported to cause immunological changes in the thymus*.

Where two words appear to have a similar structural distinction in the corpus, it can often be shown that they have varying semantic contexts, and such is the case with shown and reported. Shown is invariably followed by qualitative,

biochemical or technical explanation, while reported is associated with quantitative, empirical observations:

> The disease ⟨has been shown to be⟩ encoded by a reagent focalisation…⟨has been shown to be⟩ a prerequisite involved in the metabolism of…
> Tumor Necrosis Factor ⟨has been shown to be⟩ homologous to cachectin
> Immunological test samples ⟨have been shown to be⟩ sensitive to this process
> Lung cancer ⟨has been shown to be⟩ caused by an infectious agent
> Mutation of the p53 gene ⟨has been reported to be⟩ a very frequent event
> Transcription of cFos ⟨has been reported to be⟩ rapid and transiently enhanced
> Crystal structure ⟨has been reported to be⟩ different for a number of molecules
> Plasma levels ⟨have been reported to be⟩ both higher than (Ghanadian, 1979), similar to (Draft *et al*., 1992)…
> Associated macrophages ⟨have been reported to⟩ contain several coagulation factors

However, while shown is invariably followed by projected clauses, report is more frequently followed by prepositional adjuncts with "in" e.g. *Distinct defects have been reported in many tissues*…. There is also a large variety of research oriented verbs which introduce phrasal/prepositional complements rather than adjuncts, i.e. each verb is consistently associated with one preposition. The most frequent examples are:

Aspirin	⟨has been⟩⟨associated with⟩ gastrointestinal bleeding
Somatic mutations	⟨have been⟩ ⟨implicated in⟩ the formation of tumours
Antigens	⟨have been⟩ ⟨identified as⟩ the molecules that are expressed
A variety of mechanisms	⟨have been⟩ ⟨implicated in⟩ the development of breast cancer

The second pattern for been is less varied, has a less technical scope and involves extraposed clauses with an emphasis on signaling the general research aims of the article. The projecting verb-complement clauses in this case are past results framed in terms of a new (present tense) research direction: ⟨it has been⟩ [research process] that :

⟨it has been⟩	proposed that	this transformation involves DNA damage
	established that	they are reactive with the extracelluar domain of p185
	postulated that	the mutogenic effect of estrogens are mediated…

concluded that	MP substitution is a significant tumorigenic factor.
suggested that	thyamine is involved in the development of prostatic cancer.

The number of verbs that can be used in either pattern is restricted. Only verbs such as shown, found and demonstrated can be used in either of the two main present passive patterns (and there are very few ⟨have been demonstrated to⟩ and as we have noted there are no instances of ⟨it has been reported that⟩). The significance of this is perhaps not fully apparent until we consider that the article writers tend not to use the same forms of verbs with different tenses or aspects (*we reported that, X is/was shown to*). Similarly, to what extent can perfect passives ⟨has/have been⟩ be related to simple passives with ⟨is⟩? We find that there are just two verbs which allow projecting clauses with is ⟨is known to⟩ and ⟨is thought to⟩, usually used to introduce biochemical processes. This appears to confirm the lexico-grammatical perspective that holds that grammatical features such as aspect (here a grammatical distinction between passive perfect and simple passive) are not "free variables". Instead, there is a clear lexical correlation between clause type and the actual verb chosen, with aspect a functionally redundant feature of the lexical complex. This is further demonstrated below, in our discussion of is, to and of.

Apart from passive reporting, has/have play a key role in the phraseology of report, taxonomy and evaluation, with 46% of their combined occurrences involved in active expressions. The most frequent active expression is of the type ⟨has received⟩ where the phraseological pattern is: [clinical approach or technique] has received [some quantity of] attention/investigation followed by a reformulation of the clinical process:

> combined NMR therapy ⟨has received little investigation⟩ on a clinical basis
> PIMO antigen ⟨has received little investigation⟩ as a factor in this disease
> intracellular solvoyosis ⟨has received little attention⟩ as a possible treatment
> interferon ⟨has received much attention⟩ as a potential cure for cancer
> C1350 ⟨has received particular attention⟩ as a possible source of metabolic data.

Elsewhere in the corpus, has/have have a different set of uses. In Abstracts they usually appear in active expressions following relative clauses with a different semantic context: [patients, subjects of a clinical trial] ⟨who have received⟩ [drugs] (and also ⟨take part in⟩ [experiments]) (Gledhill, 1995b). It is interesting to note that patients are never given drugs (i.e. in a passive expression), but are expressed as actively receiving them. In Results sections, past tense had is exclusively reserved for summarizing attributes as results (*Mice had a decreased number of formation, Cells had a different correlation coefficient, animal tumours had greater mean length*). In Discussion sections has/have tend to be used in relational clauses expressing some explicit evaluation (in relational clauses the subject is either

identified or given an attribute as in: *surviving cells have aberrant morphology, the drug may have important implications, the current assays may have limited sensitivity, granisteron has been shown to have negligible agonist functions, fragments have been reported to have superior localization abilities*).

4.2. Is

As with verbs such as has/have, the distribution and use of the word "to be" is distinct within each rhetorical section of the research article. In Abstracts, the past tense is more frequent and is used to express quantitative results (*were higher, were lower than*). Was/were are also the predominant use of the verb in Methods sections, and they express passivized clinical process verbs (*were mixed, was added to*). In Discussion sections, the use of is "is" tied to the reformulation of results in projecting clauses (*the analogue is found to be a viable alternative, the effect is thought to be significant*).

In Introductions, is is almost exclusively used in relational "identifying" clauses; a rare form outside the Introductions section. Identifying clauses involve re-labeling of the subject as a new element: *X is a Y*. This is compatible with the explanatory function of Introductions sections. In science writing however, the expression *X is a Y* almost always involves explicit evaluation, combined with a reformulation of a specific biochemical process as a methodological "indicator" to be further explored:

Biochemical process		*Evaluative*	*Empirical item*
[specific disease]	is a	common	predictor
BORA	is an	appealing	alternative method
resistance to therapy	is a	critical	parameter
Pancreitis	is a	major	sign
the Winsford deposit	is an	imperfect	route

The second most frequent pattern for is in Introductions involves "attributive" clauses, where the subject is given additional attributes as part of an overall explanation. For example, only disease related items are "associated with" other disease-related cause: *toxicity, weight loss…is associated with…* Conversely, only treatment related items can be "more" [an observed property]:

target orientation	⟨is more⟩	efficient
MTX as an inhibitor		efficacious
a new foliative agent		localized
this choice of prodrug		popular
antitumour activity		stable

The reason for these patterns stems fairly straightforwardly from the research activity. Diseases are being associated with potential causes, while treatments are

being compared and measured. The phraseology is partly redundant, serving to signal and to reinforce the relationship. Seemingly intuitive, the semantics of these expressions would be far less predictable in the general language.

The collocational analysis of "is" also reveals a limited set of items which can introduce noun-predicate (projecting) clauses. The projected clause is always a biochemical fact. The projecting noun varies from empirical to research oriented terms and also usually involves explicit evaluation (here underlined). The following list exemplifies each noun we find:

disadvantage...	⟨is that⟩	a magnetic field may enhance...
The most direct evidence	⟨is that⟩	coagulation factors diffuse
A simple explanation	⟨is that⟩	none of these is currently in use
The expectation	⟨is that⟩	PTC apparently does not show mutagenesis
An intriguing observation	⟨is that⟩	these compounds are t-promoters
A major obstacle	⟨is that⟩	they repel.
An interesting outcome...	⟨is that⟩	the polar effect is masked

Among the varied projecting clause-types that Halliday (1985) identifies, this form is simply termed a "fact", where the fact-clause is labeled by the initial noun. These resemble noun-complement clauses which involve a noun + clause as in *the hypothesis that, the requirement that, the suggestion that*. The difference in function is that these clauses allow for new information to follow, but they do not involve the explicit evaluation of fact-clauses. Introductions have the widest variety of nouns involved in this structure (around 20) while Discussions have only one frequent form. We find just two evaluative expressions in Discussions: *our previous contention that, the criticism that* but these are overshadowed by the neutral ⟨the fact that⟩ (over 30) which is, by contrast, totally absent from Introductions. It is clear that nominal expressions of evaluation and conviction are more acceptable in Introductions than in Discussions. This suggests that lexical variety within a grammatical structure may be a consistent feature of rhetorical sections rather than the genre as a whole, and that overall syntactic analyses (which simply find the syntactic properties of the text or differentiate sections by syntactic arguments alone) may overlook these patterns.

4.3. Of

Of is the second most frequent item in the corpus, with 4.3% of the total number of words (this figure is 3% in a general language corpus, c.f. Sinclair, 1991). Of represents the complex nominal nature of science writing. Sinclair (1991) has pointed out its specific grammatical nature in the general language: it is not a typical preposition. In RA Introductions, of serves to qualify empirical process nouns (e.g. *characterization of...measurement of...*) and to form fixed biochemical or clinical terminology. While of is salient in Titles and Abstracts, fixed expressions and collocations (such as *effects of treatment Y*) are repeated but

also expanded to longer stretches of phraseology in Introductions. The following left/right collocates demonstrate the variety of collocation, in order of frequency:

> Left collocates (frequency >10): effects, concentration, treatment, effect, number, presence, variety, activity, results, mechanism, administration, use, because, levels.
> Right collocates (frequency >10): this, these, cells, human, compounds, drug, mice, drugs, mouse, methylene, studies, cancer, Bora, liver, cell, chloride, effects.

Although the expression ⟨effect of/effects of⟩ [treatment X] on [disease Y] dominates the use of of in all sections, the context of the word varies quite significantly from one section to another. In Titles un-premodified research-oriented items are prevalent as in ⟨Evaluation of⟩ (monitor/test) (+ing), ⟨Treatment of⟩ and the more idiomatic ⟨A case of⟩ [biochemical process] + relative clause. In Abstracts the tendency is to express empirical measurement (*a number of, a group of, frequency of, concentration of, incidence of, levels of*).

In Introductions, the most frequent use of of is in biochemical processes (*expression of, exposure of, activity of, formation of, hydrolysis of*). Of is also used in fixed expressions serving as discourse signals, the two most frequent being ⟨in view of⟩ [previous finding] and ⟨because of⟩ [presence/absence] (*In view of its oestrogen dependence, in view of our recent findings, in view of these limitations…because of the presence of hormone, because of the apparent lack of cross resistance*) and in comparisons involving the frequent collocation ⟨that of⟩ (*superior to that of the substrate, superior to that of oxygen, faster than that of the acetal function…*). Of also appears in more research-oriented phrases (the aim/purpose of this study/this report) and in nominals where the left item can be "applied" to the right item (i.e. of introduces a complement of the left collocate). One particularly interesting example of the latter is the fixed term ⟨in the treatment of⟩. It appears to be used in the reformulation of similar concepts as new drugs in a relatively long phraseology: [treatment X/new drug] (is) ⟨(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
> that Harris et al. report to be the drug of potential value ⟨used in the treatment of⟩…tumours.

In other cases, of is used as "support", i.e. in a combined lexical item that largely pre-modifies the right collocate. The most frequent is ⟨in a variety of⟩ (as

in *Enzymes are involved in a variety of anticancer drugs*). Around a quarter of these expressions involve the perhaps redundant collocation "in a wide variety of", an equally common feature of the general language. ⟨In a variety of⟩ is only typical of Introductions, the alternative expression ⟨in a number of⟩ apparently replacing it elsewhere in the corpus. Finally, of is also used throughout the corpus in the creation of complex nominals which appear to be more fixed than simple complement/modifier expressions, for example ⟨loss of heterozygosity⟩ which is terminologized and written LOH. Another example, ⟨mechanism of action⟩ is very common and is almost always mentioned as the last research aim in Introductions. We have also noted it frequently in popularized science and in other languages (Gledhill, 1997). The expression also appears to occur in a very delimited phraseological context: ⟨The mechanism of action of⟩ (disease Y/model) which is then followed by hedged or negative research process:

> ⟨The mechanism of action⟩ of human tumour model systems is
> ⟨The mechanism of action of⟩ their cytostatic action appears to be mutogenic
> Thus ⟨mechanism of action of⟩ human tumor models has not been determined with certainty
> ⟨The mechanism of action of⟩ methylene chloride has not been clarified
> However ⟨the mechanism of action of⟩ these tumor models can be deciphered
> Although ⟨the mechanism of action of⟩ some carcinogens remains unknown...

4.4. To

We have already seen the role of to in projecting expressions such as ⟨have been shown to⟩. This does not, however, exhaust its role in noun phrase complements in other salient expressions. One particularly regular projecting clause takes the following form: [biochemical process] (possessive) ⟨ability to⟩ [biochemical process]:

[the reactant] Its	ability to	alter tolerance to self
we extended its [tumor]	ability to	differentiate
calibrating their [leukocytes]	ability to	modify factor specific DNA
exemplified by its [Xpa3]	ability to	undergo epoxidation

We also find mental research processes projecting explanatory clauses:'

cells	are	⟨known to⟩	bind p53
chemicals			cause embryo toxicity
enzymes			inhibit hepatic MFO activity
hydrolysis	is		proceed via a 2-step reaction
proteins are			repair the 6-0 methylguanine

As a complementizer element of other verbs in Introductions, to most frequently occurs in ⟨appears to⟩ and it is generally used in conjunction with a negative statement, or a statement that contradicts an accompanying clause:

> Although the regulation of MyoD1 is not fully understood, this ⟨appears to⟩ perform critical functions.
> However, the function of p52…does not ⟨appear to⟩ stimulate DNA synthesis directly. Many tumours ⟨appear to⟩ have no relation to DNT oncogenic viruses
> However, this ⟨appears to⟩ contradict some of our preliminary observations.
> It ⟨appears to⟩ be an ubiquitous protein, although there is no correlation…

The phraseology ⟨appears to⟩ seems to be linked not solely with its typical role of "hedging" an assertion, but also with signaling contradiction, tied in with negative subordinate clauses. We also note that the negative which accompanies adversatives like "Although" seems to operate in parallel with "appears that" and comes either in the main or subordinate clause: it is as if the phraseology requires a negative expression but has no preference about where it is finally expressed. Again, one explanation for this variation may be that phraseology determines what grammatical choices are available with the final "mechanism" of thematic choice and word order left to textual considerations.

More generally, we have seen that to replaces that as the most frequent complementizer in Introductions whereas elsewhere in the corpus that is more salient. That clauses typically involve extrapositon and evaluation of propositions (*it has been shown that X*). As we have seen above to clauses generally involve projecting active roles to various biochemical entities (*X has been shown to* and the examples above), and this may give Introductions a more action-oriented role.

To as complementizer accounts for half of its occurrences in Introductions whereas to is predominantly a preposition in Methods sections. Typical prepositions in the corpus become highly fixed in usage (other examples include from [biochemical locus] and at [time]) and involve very specific semantic distinctions. This is also strikingly reflected in the tense patterns of the prepositional verb ⟨lead to⟩ where the past tense is used for the research oriented pattern:

These observations	⟨led to⟩ comparative studies
these findings	⟨led to⟩ widespread use of hormonal aspects
Identification of cell response	⟨led to⟩ the investigation of radioimmunization
we describe the rationale which	⟨led to⟩ speculation that 5HT3 receptors…
These results	⟨led to⟩ the selection of a battery of immune assays

While the present tense is exclusively used for the biochemical/technical pattern:

response to DNA damage	⟨leads to⟩	an arrest of the cells
This in turn	⟨leads to⟩	increased conversion of the lactase

This process	⟨leads to⟩	inhibition of intracellular concentrations
altered membrane transport	⟨leads to⟩	degradation extracelluar matrix (ECM)
the agonist 2-methyl 5HT	⟨leads to⟩	release of substance P

Again, the rationale for this intriguing difference is that tense and aspect play a role in phraseology and that tense has a "research orientation" that is more lexically determined than dependent on a strictly grammatical category. In ESP research, there has been a tendency to see tense as a grammatical category that has validity across verb forms, for example the present is seen to express research established before the work carried out in the current article (Hanania & Akhtar, 1985), while the present perfect is reserved for the expression of current findings (Gunawardena, 1989). The analysis of phraseology we have seen above in led/lead, and even in the "grammatical" verbs have/be, adds an extra dimension to this kind of analysis: the verb form and tense are indeed consistent, but verb form and function are also very dependent on local constraints.

Our final observation of phraseology involves was (the fourth most frequent collocate of to) where almost all of the expressions formulate the aims of the research article. There is a very wide variety of expression and the main forms are listed below:

The aim of this study ⟨was to⟩ compare
The intention ⟨was to⟩ determine
One further goal ⟨was to⟩ evaluate
The key to the plan ⟨was to⟩ examine
Therefore our second objective ⟨was to⟩ expand data
their policy ⟨was to⟩ examine
Our purpose ⟨was to⟩ explore whether
Another goal of these studies ⟨was to⟩ identify DNA adducers
The aim of the present series of these studies ⟨was to⟩ investigate
The present studys aim ⟨was to⟩ investigate whether
The goal of this study ⟨was to⟩ re-evaluate
A main task ⟨was to⟩ study whether
Thus, the first aim of the present study ⟨was to⟩ test
The purpose of the Bristol 3rd stage trail ⟨was to⟩ use
The purpose of this work ⟨was to⟩ widen the research window...
(The purpose of the current report ⟨was to⟩ generate and trap...)

Syntactically these resemble the "fact" clauses we considered under "is": but here we have "act" clauses, the same projecting noun type but this time followed by a to-clause. The contrast of tense ⟨was to⟩ vs ⟨is that⟩ is consistent, and possibly purely conventional. The semantic pattern sticks very consistently: [research goal] ⟨was to⟩ [research process]. Expressions such as *the current/present report/study* are quasi-obligatory parts of the whole, and once the researcher has chosen to express the idea of an "aim" his or her next choice is limited to the form ⟨was to⟩ and

then either V(research) + phenomenon or V(research) + whether + (proposition). These expressions can involve long stretches, but their overall structure is predictable and we refer to them elsewhere as collocational cascades (Gledhill, 1995b). The principle of the cascade is that each choice is indexed to a smaller range of choices but when we arrive at a fixed expression the range of choice is expanded again.

The only exception to this seems to be where the aim is to "do something", in other words the technical clinical process generate and trap. This may seem unsurprising, but the important point about phraseology is that perfectly plausible alternatives such as generate and trap are not equally as prevalent as research process expressions: they are exceptions. There is no logical reason why the potential expression [research goal] ⟨was to⟩ [empirical/clinical process] should not occur just as frequently in the corpus (as in *Our aim was to hydrolyse drug X* or *The purpose of this report was to produce a solution of chemical Y*). It is plausible to imagine that this is a function of the expectations of the discourse community of medical researchers, as opposed to biochemists in general. A corollary is that what would be free or restricted collocation in the general language becomes fixed in the specific language. In the case of stating aims, it is clearly preferred that goals be presented as global research rather than the specific empirical or clinical processes.

5. Conclusion

> The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments. To some extent, this may reflect the recurrence of similar situations in human affairs; it may illustrate a natural tendency to economy of effort or it may be motivated in part by the exigencies of real-time speech. (Sinclair, 1987: 320)

We hope that we have been able to demonstrate the idiom principle at work in science writing. In some instances collocation involves terminology and reflects the recurrent semantics of the specialist domain. In other instances collocation reveals the dominant discourse strategies in the research article. We can consider that both levels are integrated into the intermediate level of phraseology, or "the preferred way of saying things". Our examination of the pharmaceutical sciences corpus points to lexico-grammatical correspondences that are particular to the cancer research article genre and it should now be possible to observe different lexico-grammatical correspondences at different time periods (i.e. in the evolution of the research article genre) and at different levels of specialization (i.e. in the popularization of science). These are areas that corpus linguistics is well placed to exploit. In addition, the direct correspondence between lexis and grammar is now

so pervasive that it is difficult to conceive of a general characterization of science writing or the design of teaching materials for the benefit of science writers which can afford to ignore phraseology as a central level of analysis.

To what extent do the patterns we have observed above have a role to play in the discourse of cancer research? There appear to be a number of implications:

1. Communicative competence in the LSP includes a subconscious knowledge of collocations.
2. Collocation allows for predictability and contrastive innovation within the text.
3. Phraseology is part of the defining characteristics of the discourse community.

It is clear that such regular phraseology cannot have emerged from the preferences of a small set of editors or the practices of one particular journal. None of our expert informants were able to recall any editorial policy that prohibits patients from *being given drugs* or requires projecting nouns clauses to be limited to specific sections of the article. Instead they attested to a general lack of training in written communication skills. It follows that the phraseological units we have identified are formulated by previous discourse and must be acquired or learnt by the community. If the collocational patterns are so idiosyncratic and yet pervasive, it is possible to conclude that the cohesive mechanics of the discourse community appear to be stronger than previously imagined, even if these are largely invisible. We might also argue that a systematic linguistic trait of phraseology reveals an orientation which is deeply rooted in the ideology of the discourse community. In the expression "*patients who received drug X*", science consistently considers patients and subjects as agents, concomitant with the distancing of the researchers from any clinical process. This would correspond to Stubb's (Stubbs, 1996) approach which attempts to find correspondences with recurrent speech styles and an ideological stance in language.

Our conclusion is that collocational patterns indicate a wider relationship beyond the individual text and reflect an evolutionary process that has forged the conventions of a number of phrases in the language of cancer research. There is now a body of linguistic theory that sees lexico-grammatical patterns as central to the way discourse is *construed* (Halliday & Martin, 1993), how we build and interpret the world though discourse. This neo-Firthian view of language sees the semantics of the word as textually distributed, and syntax as intimately linked with lexical knowledge. Myers (1991) and Hoey (1991) have noted that lexical choice in particular constrains the textual choices that the writer may make in later discourse and that the reader uses collocation in order to skim and scan across the text and to interpret new co-occurrences. The ability to interpret new items in the light of previously existing collocations has also been suggested by Pavel in the field of terminology:

> ...new turns of phrase generate meaning, condense into stable expressions of those meanings and become first synonymous neologisms, and then terms that give birth to new terms. (Pavel, 1993)

This view of language promotes the probabilistic nature of variation, where variation is seen as a product of a series of selections which themselves affect the probability of future selections of expression (Halliday & Martin, 1993). How do scientists know which collocations bear at a point in the text? The processes are unclear, but Halliday and Martin claim that instantial knowledge is an important concept in understanding textual development. Instantial knowledge is determined at the point of expression in the text, it is the kind of knowledge that allows us to interpret the meaning of a new term based on our reading of the text rather than our background knowledge. The new term's associations and extensions will be built up by the text. Instantial knowledge affects, for instance, which tense to use in expressing biochemical and research processes, which valences to adopt when in relative clauses and so on. These decisions are in part phraseological, and in part textually determined. However, in keeping with Halliday's view of recurrent selection, each textual decision will stand a good chance of affecting later decisions in other texts. Instantial knowledge can then be seen as a central factor in the process of writing and reading in this specialist field and in the creation, maintenance and reformulation of phraseology.

But it is equally important not to impose an interpretation on collocations as having a fixed function. Although there may be a good temporal explanation of why "leads to" and "led" to have differing phraseologies, or why "in vitro" and "in vivo" have different syntactic functions, it may also be that the original function has been lost. Many aspects of these collocations are functionally redundant, that is to say that their function has moved from a literal meaning to broadly conventional practice that exists to allow readers to predict relations within the text. This may account for the large scale redundancy that Banks (1994b) suggested in research article Discussions.

Finally, we can conceive of collocations as cultural pieces of information, Dawkins (1976) "memes", transmitted from one researcher to the next. Whatever the function of collocations, their mere existence suggests that expressions are replicated by successive generations of science writers. The concept of cultural evolution through language has not escaped certain researchers in cognitive linguistics recently:

> Languages are inanimate artifacts, patterns of sounds and scribblings on clay or paper, that happen to get insinuated into the activities of human brains which replicate their parts, assemble them into systems, and pass them on. The fact that the replicated information that constitutes a language is not organized into an animate being in no way excludes it from being an integrated adaptive entity evolving with respect to human hosts. (Deacon, 1997:112)

"Memes" take the form of any small cultural entity that can be remembered as a unit such as a snatch of song, a recipe, a proverb. We would argue that just like memes, collocations can suggest larger units and they are usually transmitted whole from one speaker to the next. Just like genes, collocations may not serve their original function and in the form of idioms their structure may not

correspond to their meaning in the text. Collocations in science writing are undoubtedly selected as the best ways of expressing certain ideas, although this selection does not mean that these expressions are the best, or the only possible selections; the selection is largely a feature of convention and acceptability within the discourse community. There are many instances in science writing where breaking with the phraseological conventions has been seen to be part of the process of change and innovation, as well as acceptance within the discourse community as Myers (1990) demonstrated. Scientists are not good at identifying collocations themselves, but they are aware of ''catch phrases'' and ''good academic style'' which must in the end be realizable as collocations. It is clear that collocations are part of the mechanisms by which specialist writing is internally cohesive and by which it evolves. An understanding of the adaptive processes in written science, in particular the processes that forge systematic, recurrent examples of written language is far from complete. The discourse processes behind collocation are likely to remain unclear, but we hope to have demonstrated that corpus analysis is at least the first step in the process of building a new model of language that at once takes account of the lexico-grammar but which also leads to a deeper understanding of the obligatory nature of much that is written in academic discourse.

Acknowledgements

I would like to thank Dr Mike Scott at Liverpool for his time and the invaluable use of his computer programs. I am also particularly grateful to all the authors and publishers who allowed for the creation of the Pharmaceutical Sciences Corpus.

Appendix 1

1.1. Constitution of the Pharmaceutical Sciences Corpus (By Articles Included)

International Journal of Cancer 25

Cancer Chemotherapy and Pharmacology 16

Cancer Research 12

British Journal of Chemistry 11

Journal of Organic and Applied Chemical Studies 11

Carcinogenesis 10

Fundamental and Applied Toxicology 10

Journal of Chemistry Perkin Transactions 10

Cancer Letters 9

Journal of General Microbiology 9

Journal of Chemistry 7

British Medical Journal 5

British Journal of Pharmacology 3

References

Aijmer, K., & Altenberg, B. (1991). *English corpus linguistics*. London: Longman.
Altenberg, B., & Eeg-Olofsson, M. (1990). Phraseology in spoken English: presentation of a project. In J. Aarts, & W. Meijs, *Theory and practice in corpus linguistics*. Amsterdam: Rodopi.
Banks, D. (1994a). Clause organization in the scientific journal article. *ALSED-LSP Newsletter*, *17*, 4–16.
Banks, D. (1994b). *Writ in water: aspects of the scientific journal article*. E.R.L.A.: Université de Bretagne.
Barnbrook, G. (1996). *Language and computers*. Edinburgh: Edinburgh University Press.
Biber, D. (1986). *Variation across speech and writing*. Cambridge: Cambridge University Press.
Biber, D., & Finegan, E. (1994). *Sociolinguistic perspectives on register*. Oxford: Oxford University Press.
Bittencourt dos Santos, M. (1996). The textual organization of research article abstracts in applied linguistics. *Text*, *16*, 481–500.
Cruse, R. A. (1984). *Lexical semantics*. Cambridge: Cambridge University Press.
Dawkins R. (1976). *The Selfish Gene*. Oxford: Oxford University Press.
Deacon, T. (1997). *The symbolic species: the co-evolution of language and the human brain*. London: Allan Lane—The Penguin Press.
Gledhill, C. (1995a). Collocation and genre analysis. The discourse function of collocation in cancer research abstracts and articles. *Zeitschrift für Anglistik und Amerikanistik*. *1*, 1–26.
Gledhill, C. (1995b). *Scientific innovation and the phraseology of rhetoric*. Unpublished PhD thesis. Aston University, Birmingham.
Gledhill, C. (1996). Science as a collocation. Phraseology in cancer research articles. In S. Botley, J. Glass, T. McEnery, & A. Wilson, *Teaching and language corpora* Vol. 9, pp. 108–126. UCREL Technical Papers.
Gledhill, C. (1997). Les collocations dans la construction du savoir scientifique. *ASp, Groupe détudes et de recherche sur langlais de spécialité*, *15–18*, 85–104.
Gunawardena, C. N. (1989). The present perfect in the rhetorical divisions of biology and biochemistry journal articles. *English for Specific Purposes*, *8*, 265–273.
Halliday, M. A. K. (1985). *Introduction to functional grammar*. London: Edward Arnold.
Halliday, M. A. K., & Hasan R. (1976). *Cohesion in English*. London: Longman.
Halliday, M. A. K., & Martin J. (1993). *Writing science: literacy and discursive power*. London: Falmer Press.
Hanania, E. A. S., & Akhtar K. (1985). Verb form and rhetorical function in science writing: a study of MSc theses in Biology, Chemistry, and Physics. *English for Specific Purposes*, *4*, 49–58.
Hoey, M. (1991). *Patterns of lexis in text*. Oxford: Oxford University Press.

McEnery, T., & Wilson, A. (1996). *Corpus linguistics*. Edinburgh: Edinburgh University Press.
Moon, R. (1992). There is reason in the roasting of eggs. A comparison of fixed expressions in native speaker dictionaries. *Euralex 92: proceedings* (pp. 493–502). Oxford University Press.
Myers, G. (1989). The pragmatics of politeness in scientific articles. *Applied Linguistics*, *10*, 1–35.
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. *Discourse Processes*, *14*, 1–26.
Pavel, S. (1993). Neology and phraseology as terminology-in-the-making. In H. B. Sonneveld, & K. L. Loening, *Terminology: applications in interdisciplinary communication* (pp. 21–34). Amsterdam: John Benjamins.
Renouf, A., & Sinclair, J. McH. (1991). Collocational frameworks in English. In K. Aijmer, & B. Altenberg, *English corpus linguistics* (pp. 128–144). London: Longman.
Sager, J. C., Dungworth, D., & McDonald, P. F. (1980). *English special languages: principles and practice in science and technology*. Wiesbaden: Oscar Nadstetter Verlag.
Salager-Meyer, F. (1992). A text-type and move analysis study of verb tense and modality distribution in medical English abstracts. *English for Specific Purposes*, *11*, 93–114.
Sinclair, J. McH. (1987). *Looking up: an account of the Collins cobuild project*. London: Collins ELT.
Sinclair, J. McH. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Stubbs, M. (1996). *Text and corpus analysis*. London: Routledge.
Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
Thetela, P. (1997). Evaluated entities and parameters of value in academic research articles. *English for Specific Purposes*, *16*, 101–118.
Thomas, P. (1993). Choosing headwords from LSP collocations for entry into a terminology data bank (term bank). In H. B. Sonneveld, & K. L. Loening., *Terminology: applications in interdisciplinary communication* (pp. 46–68). Amsterdam: John Benjamins.

Chris Gledhill is lecturer in the French Department at the University of St Andrews. He taught Advanced TEFL, Linguistics and French at Aston University, Birmingham and completed his thesis there in 1995 (Scientific Innovation and the Phraseology of Rhetoric). His research involves comparative projects phraseology in science writing, artificial languages and French.