Collocation and genre analysis. The Phraseology of grammatical items in cancer research articles and abstracts

Christopher Gledhill

Gledhill, Christopher. 1995. Collocation and genre analysis. The Phraseology of grammatical items in cancer research abstracts and articles. Zeitschrift für Anglistik und Amerikanistik 43 (1/1). 11-36. ISSN 0044-2305.

Chris Gledhill

Collocation and Genre analysis. The Phraseology of grammatical items in cancer research articles.

This paper presents a systemic view of collocation in order to explore how scientists optimally select and present information in abstracts. In a previous large scale study of 150 research articles (Gledhill 1995), collocational analysis revealed the key role of idiom in defining a norm of scientific writing and in the progressive textual creation of new scientific ideas (associated with the concept of logogenesis: Halliday and Martin (1993)). In order to deal with the analysis of textual genres it is necessary to conduct corpus analysis that is sensitive to the professional goals and practices in which a specific text type is couched. This article describes an approach to corpus analysis which marries the context-sensitive approach of genre analysis (Swales 1990) with the latest computational techniques for establishing global linguistic properties of text (Sinclair 1993). In particular, the notion of phraseology can be seen to operate at a level that incorporates the collocational properties of high frequency items with what has come to be known as 'terse text'.

Introduction

This article attempts to examine the lexico-grammar of cancer research abstracts, and argues that corpus linguistics can usefully address the concerns of the discourse analyst and student of English for Specific Purposes. Studies of abstracts are prevalent in the information sciences, as we mention in section 1 below, but there have been fewer linguistic analyses. In the specific field of genre analysis there have been few studies of abstracts, and no studies we know of that exploit the notion of phraseology. Abstracts written by authors have been characterised in terms of morpho-syntactic features, especially verb tense and modality (Hanania and Akhtar 1985, Malcolm 1987, Gunawardena 1989, Salager-Meyer 1992). Discourse analysis has involved comparison of rhetorical moves between abstracts and articles (Nwogu 1989, Endres-Niggemeyer 1985, Salager-Meyer 1992) and thematic choice between successful and non-successful abstracts (Nwogu and Bloor 1991, Drury 1991, Gibson 1992).

Generally, abstracts appear to have been analysed in terms of 'register analysis', that is as a text-type defined by its shared lexico-grammatical features (Halliday and Hasan 1976). Since a fundamental assumption of the register approach is that "aspects of

context and features of language can be aligned" (Hunston 1995:13), the same linguistic features are seen to express similar functions across text-types. This has allowed Halliday to argue that scientific texts create textual meaning using linguistic mechanisms such as grammatical metaphor (Halliday and Martin 1993). Increasingly, register studies depend on computer corpora that are grammatically marked up or 'tagged'. Kretzenbacher (1990) and Atkinson (1992) use such a methodology in their characterisation of academic abstracts and articles. Biber and Finegan (1986-) have been primary exponents of computer-based register analysis, an approach that measures variation in texts by the occurence of linguistc features. They identify dimensions such as 'abstractness' and 'explicit information' that emerge from the co-occurrence of grammatical features such as clause complexes, it-clefts, adverbials, and more recently, lexical chains and deictic anaphora (Biber 1992). Essentially, they maintain that text-types exist on a continuum, and that textual differences can be explained by internal linguistic properties of texts as opposed to external social and rhetorical features.

The 'genre' approach on the other hand offers an alternative perspective. It assumes that the professional goals (Swales 1990) or a broader more embracing concept of ideology (Martin 1985) define the text-type. According to this approach, functions of linguistic features change for each genre and discourse community. However, while gener analysis maintains a Some features of rhetorical structure, such as explicit discourse signals, have already been identified in the genre analysis of scientific articles alone (Oster 1981, Tadros 1985, Master 1987, Brett 1994) and to a lesser extent in abstracts (Diodato 1982, Zambrano 1987, Krezenbacher 1990). To my knowledge, however, corpus analysis has not as yet been exploited from a genre analysis perspective, and there has not been a general phraseological comparison of research articles and abstracts.

A phraseological approach (as outlined by Moon 1992, Francis 1993) attempts a global desecription of the lexicogrammar as it corresponds to the rhetorical features of a representative corpus. Phraseology is a system of preferred expressions differentiated by the rhetorical aims of a discourse community. The concept addresses both issues of lexis and discourse. In terms of lexis, the concept of collocation has been used in the analysis of the intermediate level of language between syntax and lexis. Recurrent word patterns have also been instrumental in recent developments in

lexicography and the description of English, as in the Cobuild project (Sinclair 1987, Francis 1993). On the level of discourse, phraseology plays an important role in rhetorical choice, and idioms have been claimed to constitute important stages in the rhetorical development of texts (Moon 1992, McCarthy and Carter 1993). Pavel (1993) presents a textual view of phraseology, tracing the development of collocations within texts where deviation from the norm implies innovation and neology in the scientitific community. This aspect of textual development touches on the concept of logogenesis, which is the subject of ongoing research (Halliday and Martin 1993, Hunston 1995, Gledhill 1995a).

Computer-based corpus analysis of the kind practised by the Cobuild and Longmans dictionaries has advanced the analysis of phraseology and emphasised the role of lexis as opposed to syntax. But the corpus analysis of phraseology has not been exploited in terms of more specialised language. A secondary aim of this paper is to contextualise corpus analysis by moving away from analysis of registers to contextually defined genres. The linguistic features discussed here are to be seen as unique constructs, exploited in a unique socio-cultural setting (in this case research articles published for an academic community of scientists in the field of cancer research). This first involves setting out the context of use of a particular genre and secondly analysing a representative corpus of the genre in collaboration with its users. We turn first to the role of abstracts in science.

1. The scientific abstract.

Before embarking on a discussion of the lexico-grammar of abstracts, the current context of the abstract in science writing needs to be considerd, emphasising especially the unexpected complexity of a genre that is often considered to be a cut and dry product (or worse, an *extract*) derived from an original text (Maizell et al. 1971, Cremmins 1982, Cleveland and Cleveland 1983, Bernier 1985 inter alia).

In the complex world of scientific communication, the research article and abstract in a refereed journal are considered to be key marketing tools not only in the raw give-and-take of specialised facts, but also in the maintenance of a community hierarchy (Knorr-Cetina 1983:106, Swales 1990). For many reasons, linguists have been attracted to the study of summaries and similar texts, since they represent a

highly crafted text type where economy of expression and coherent presentation of ideas are vital. At the same time the abstract in the experimental sciences is essentially what Lane (1992) terms a 'péritexte', a disembodied and self-standing reference tool, where there appear to be two functions from the perspective of the discourse participants:

1- The author aims to propagate his or her research, either by successfully pitching the message or by convincing the reader to read on.

2- The reader aims to assimilate information by either avoiding lengthy reading of the main article or by efficiently accessing the information in the research article, bearing the contents of the abstract in mind.

The field of English for Specific Purposes (ESP) has particularly had to come to terms with the genre, given the large number of non-native speakers of English who need language training in this area. (Gibson (1992) Endres-Niggemeyer (1990,1991)). However, with the advent of 'hybrid' texts (Schäffner and Adab forthcoming), textual genres can be seen to merge or develop in unexpected directions and this had meant that ESP cannot reply on one model such as rhetorical structure as a hard and replicable training methodology. As Swales notes, the research article itself is not equivalent across research disciplines and places of publication.

The research article and its abstract are generally accepted as the 'favoured' (although not the only) genre in a complex system of scientific communication. The format, and therefore the language of the genre have become a matter of convention. As Atkinson (1992) has demonstrated, the conventional rhetorical sections of scientific articles that we have come to expect (Introduction, Methods, Results and Discussion - the *IMRD* structure) are in fact recent developments of a longer process of gradual textual evolution. The culture of empirical science has moved from producing informal, narratives, aimed at replication by essentially like-minded colleagues, to formal context-dependent argumentations that are intended to attract debate and attention within a hierarchical and funding-conscious discourse community (Swales 1990). As a register, where the linguistic properties are said to embody (or 'construe') the context (Halliday and Martin 1993), there has been an evolution largely from reporting to expository research articles. As Halliday and Martin (ibid) argue, this is bound to be accompanied by an evolution in the lexico-grammar of science and in particular, the use of grammatical metaphor.

The individual roles of the title and abstract also point to the fluidity of genres. In the light of scientists' increased use of electronic indexing journals, disembodied informative abstracts written by expert indexers appear to take on the essential informative functions of the original research article (Hutchins 1987). At the same time, titles may be taking on the role of abstracts in that results are explicitly stated. This is reflected in the increasing usage of titles with finite clauses in the newer disciplines such as developmental biology (Jaime-Sisó 1993). Halliday and Martin (1993:17) argue that such congruent finite expression is more flexible in reporting research processes, although Hunston (1995) has arguued that a nominal is just as useful for reporting results (for example: *Failure of compound X to do Y*). I shall attempt to demonstrate below that grammatical features have very specific functions in specialist discourse, and that it may be oversimplisitc to assume that a feature such as grammatical metaphor operates in the same way across genres.

Returning to the theme of phylogenesis, or developing genres, it appears that there is a process of 'miniturisation' taking place analagous to the process of industrial mass production and this will inevitably have consequences on the information load of abstracts and other 'terse texts'. This changing use of information can be seen in the 100% increase in the number of indexed abstracts produced from 1969 to 1990 to over 500 000 (*Chemical Abstracts Service* figures in Maizell et al. 1971 and Metanomski 1991). This has not been met by an equally increasing number of journals, although the exponential rise in their number is likely to contribute to the overall figures. It is likely that the increasing number of abstracts reflects the correspondingly diverse growth of research outlets, with new patent journals, on-line data bases, grant proposals published on the internet, electronic journals, conferences and even abstracting indexes and PhDs on the web. The disembodied abstracts of the paper-based indexes and on-line data bases have themselves been supplemented by a natural language indexing system *Permuterm* (SCI 1993) which searches references for combinations of terms that authors use rather than terms used by indexers. Relevance is no longer judged prelimiary reading but on the judicious choice of of indexing terms. Importantly, indexing terms are no longer just nouns indicating specialist field or specific entities; they include stereotypical phrases or 'permuterms' and can thus include complex nominals and limited finite expressions (SCI 1993). The linguistic analysis of abstracts needs to reflect the technical alternatives available

and the information retrieval processes which are typically used by researchers.

Using a longitudinal ethnographic approach of observation outlined by Myers (1990), I interviewed fifteen cancer researchers working for Aston University's Pharmaceutical Sciences Depertment (Gledhill 1995a). I found that their research articles are crafted specifically for use in the lab. In particular research articles in the fields of organic chemistry display non-linear, indexical mechanisms that allow readers to consult specific parts of the text. Biochemists skip whole sections of research articles, enter the text from reference points under diagrams, and refer indexically within the text to coded 'synthetic stories' represented by chemical reaction diagrams (JOC 1993). More importantly, researchers are able to make informed guesses about the contents of the research article on the basis of field-specific information they have partially gleaned from the title or abstract. In addition, visual clues are an important process which may affect fundamental linguistic systems. For example, pronouns become redundant when chemical compounds are referred to by instantial, text-dependent identity codes. Keywords are important clues to whether a researcher will read on, but the researcher also comes to the text with pre-framed questions such as "Is this of marginal curiosity or general interest?", "Are there new or surprising findings here?", "Is there enough evidence?" and "Do I believe this?". If a researcher decides that there is not enough evidence, then the rest of the search for information is already directed along a certain path. If the researcher has already decided that the information is marginal, then no more effort is spent - a process of homing in which has been examined by Nystrand (1986).

So, as the researcher trawls data-bases, skims through abstracts and acts on key-phrases to identify central sections of texts, it appears that often neither the article nor the abstract is necessarily the first point of contact between the researcher and the information in its most accessible form. Importantly, the coherence inferable in the text is dependent on the researchers' experience and motivation for reading: thus the research article is not just a one-process text to be decoded from top to bottom, but responds (even if unintentionally) to a range of possible uses and users. I stress here the non-linearity of the scientific abstract and article, because this has to be taken into account when we consider the phraseoloigcal patterns that occur in the abstracts corpus.

2. Corpus collection: the Pharmaceutical Sciences Corpus

In this article we limit our analysis to typical writing strategies in cancer research abstracts. For this purpose, 150 papers (500 000 words) were collected with the collaboration of 15 expert informants from Aston University's Pharmaceutical Sciences Department. The papers were scanned by an electronic optical reader and placed on a IBM PC hard disk for automatic analysis. Particular emphasis was given to the design of the corpus. The corpus had to represent papers from the producers' and well as the recipients' point of view, and the corpus was designed to represent not just the output of the Pharmaceutical Sciences Corpus (PSC), but also the general field of cancer research. This turned out to be a complicated set of criteria. Sinclair (1991) has set out guidelines for the elaboration of a computer-held corpus for linguistic analysis. He emphasises the need for the corpus analyst to signal his or her motivation, his or her criteria for text selection, the internal constitution of the corpus and (as Atkins et al. 1992 put it) its imbalances: an exercise which necessarily involves input from expert informants in the Pharmaceutical Sciences Department, an essential component of genre analysis set out by Selinker et al. (1981).

Unlike the text collections used by Biber and others, the corpus described in this paper has been selected to reflect a sample of two subgenres (abstracts and articles) reflecting a specific global research aim (cancer research) in consultation with the authors and users of the articles concerned. Fifteen research academics were contacted as expert informants (Gledhill, forthcoming). In an adaption of Swales' (1990) definition of members of a discourse community, they are 'associate' members, in that while they all have different specialisms ranging from the description of the disease to the analysis of organic drugs, but they all justified their research in terms of the search for a cure for cancer and similar viral diseases: their means differ, but their ends are ultimately the same. This is of course complicated by the fact that cancer itself is not a simple entity nor a singular problem and different researchers had a different perspective on this. Cancer energed from the survey as a distributed concept: a genetic disease explained as a multiplicity of internal cellular triggering mechanisms and external metabolic causes.

In order to build the corpus, researchers were asked to:

1- donate papers they had published (10 texts)
2- recommend high prestige journals (80 texts, taken randomly from the selection)
3- recommended papers that were central to their own work. (36 texts)
4- suggest journals in order to make the corpus representative of the field of cancer research rather than the specialisms of the department. (44 texts largely from *ADONIS*)

Many texts could be down-loaded copy-right free electronic indexing services. Papers from charity journals are also copy-right free (Gledhill 1995a). Having obtained copyright permission for the rest, all 150 research articles were electronically copied (scanned) and once on disk, split into different rhetorical sections or 'subgenres' (title, abstract, introduction, methods, results, discussion). Here is a breakdown of both of these together with the vital statistics of the corpus (a complete breakdown, journal by journal, is available in Gledhill 1995):

Total number of Journals: 22.
Total number of research articles: 150.
Total number of words: 500 000 (including Exerimental and joint Results-Discussion sections)

Table 1: Subgenres in the Pharmaceutical Sciences Corpus

Title (150)		2 123		0.4%
Abstract (150)		29 136		5.7%
Introduction (150)		60 809		11.7%
Methods (125)		113 089	23.5%	
Results (120)		123 084	25.7%	
Discussion (125)	114 205	23.9%		
Total (TAIMRD only)		*443 735*	*100*	

The format of certain genres was complicated largely by the addition of experimental sections and the replacement of separate results and discussion sections by joint RD sections. This should not concern us here, since we are concentrating on how the abstract differs lexically form the rest of the article as a whole. The breakdown per journal is as follows:

Table:2 Distribution of Research articles in the PSC Corpus.

Journal.	
International Journal of Cancer	25
Cancer Chemotherapy and Pharmacology	16
Cancer Research	12
Journal of the American Chemical Society	11
British Journal of Cancer	11
Fundamental and Applied Toxicology	10
Carcinogenesis	10
Journal of the Chemical Society (Perkin Trans.)	10

Cancer Letters	9
Journal of General Microbiology	9
Journal of Organic Chemistry	7
British Medical Journal	5
British Journal of Pharmacology	3
Journal of Pharmacy and Pharmacology	3
Pharmaceutica Acta Helvetica	2

Chemical Communications, Biochemistry Journal, Tetrahedron Letters, Trends in Pharmaceutical Sciences, Journal of Medicinal Chemistry, Journal of the National Cancer Institute, Angewandte Chemie: : 1 each

Usually researchers' own submissions account for the under-represented journals and one-offs (such as Angewandte Chemie). Once the texts were scanned and post-edited, they were stored in a UNIX database, and transferred to personal computer for collocational analysis with the *Microconcord* program and for comparative frequency analysis with the *Wordlist* program (Scott 1993). While Scott has developed the *Wordlist* program to identify keywords between different texts with a view to automatic analysis of genre, the program has also proven particularly useful in the analysis of different rhetorical sections within the same texts.

3 Corpus Analysis: the typical phraseology of abstracts.

The analysis we set out here attempts to demonstrate certain key phrases that appear more frequently in the cancer research abstract than in the article. For the purposes of this paper a straight comparison between the abstract subcorpus and the main PSC corpus should provide evidence of the varying phraseology of certain key words in the abstract. The Cobuild corpus of 17 million words is used here as an inital comparative frequency list (for details c.f. Sinclair) A comparative frequency list analysis with the Cobuild corpus of a broader selection of the language (Sinclair 1987) reveals interesting patterns, even in the first ten most frequent words of the corpus, as the table below shows:

Table 3: The Wordlist top ten lexical 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

The prevalence of prepositions in the PSC corpus (except for *to*) may indicate interesting first analysis of the phraseology of these items: it would certainly help to determine which phraseology is genre dependent and which belongs to wider areas of language use. For example, *was* is usually associated with passive expressions of technical verbs in the methodology section: *the mixture was homogenized, a trocar was protonated* and so on. So the presence of *was* and *were* above indicates that this methods-oriented phraseology is typical of the PSC corpus, and atypical of the general types of language included in the Cobuild corpus. It is assumed that salient phraseology that emerges in the frequency list of all the abstracts in the corpus when copmpared with the PSC corpus as a whole should capture the typical use of these items in the abstract. When an item is more frequent in one rhetorical section than another, we shall use the term 'salient item' to indicate this. Thus below we analyse the first ten salient grammatical items in cancer research abstracts.

The program *Wordlist* makes two frequency lists, in this case a frequency list of the abstract sub-corpus (29 203 words) and the whole PSC corpus list (500 000 words) and calculates the statistical significance of each item in the abstract subcorpus (Butler 1985:176). Highly significant items appear at the top of the list (getting near to p=0.000) and this indicates that, in comparison with the rest of the corpus, the item is more likely to occur in the abstract. The items at the top of this list are abstract's 'salient' items. Words occurring at the bottom of the list are not typically involved in the abstract, and are listed in Appendix 1. The first ten 'significant' salient items (as seen in table 4 below) indicate the lexical specificity of the abstract and are often very low frequency items.

A *Wordlist* listing of all the words in a subcorpus provides us with a list of salient items that are of mixed frequency in the PSC corpus. These items could be sorted in a number of ways:

1) Highly significant lexical items.
2) Highly significant items of high frequency in the PSC corpus.

3) Highly significant grammatical items.

Statistically the PSC is too small to provide interesting phraseological data for low frequency lexical items (criterion 1) and the kind of data from a simply high frequency list (criteria 2) would be more suitable for a lexicographic or terminological survey than a phraseological one. On the other hand, few phraseological studies have concentrated on grammatical items (criterion 3), because the amounts of data to be analysed are too large. Ironically, these studies are also too large to provide insights about specific genres. The idiom principle suggests that a phraseological unit must contain at least one grammatical item. It follows that any lexical items of interest should emerge as organising elements of phraseology without us having to elicit them. In other words an analysis of phraseological structure minimises the amount of data needed by characterising global patterns first. Thus we argue that grammatical items give the optimum amount of phraseological information for a medium-to-small sized specific corpus such as the PSC.

The salient grammatical items for the six main rhetorical sections in the corpus are listed in the table below. For comparison, grammatical items that are more frequenct in the Cobuild corpus than in the PSC corpus, are in **bold**:

Table 4: Salient grammatical items in the PSC rhetorical sections.

TITLE	*ABS*	*INTRO*	*METHODS*	*RESULTS*	*DISCUSSION*
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 by	was	**can**	each	after	in
7 via	**that**	**it**	and	**there**	**not**
8 with	**did**	**we**	from	**the**	**this**
9 -	**who**	of	after	**when**	**we**
10 -	both	**to**	with	**all**	**have**

Only eight grammatical items are statistically significant in Titles. As might be expected, some sections are more 'Cobuild-like' than others. It is perhaps strange that

29 of the 58 items we analyse are more typical of cobuild than of the PSC corpus itself. The difference is that patterns attributed to Cobuild-salient items may represent a 'general language' quality of that rhetorical section. PSC-salient items on the other hand would have patterns which indicate that a particular pattern has moved the corpus as a whole away from the general language. In other words, all the grammatical items we analyse characterise a particularity of the rhetorical section that sets it apart from other sections. This kind of measure allows us to claim that we are analysing the prototypical structures rather than simply idiosyncratic features.

Francis (1993) has argued that there are strong lexical constraints on syntax, and posits that the the distinction between 'closed' or 'open' class items is blurred: we shall refer instead to lexical and grammatical as extremes in a cline (Halliday 1985). However, while lexical items can be informally associated with the knowledge structure of the research activity and the topic of cancer research, high frequency items such as the salient grammatical items we analyse below, have been shown to capture lexical phraseological patterns that represent the favoured way of presenting and reformulating information in a journalistic genre (Gledhill 1994). In order to demonstrate this and discern the most salient features of the abstract, the grammatical items of the abstracts subcorpus are isolated, and a concordance is created for each one (an example concordance of the word *of* is presented in Appendix 4). While the concordance often reveals a great deal of data as visible patterns, collocational patterns can be calculated to back up intuitive remarks, although it is often useful to remember that collocational patterns cannot pick up large scale patterns of low-frequency words which need to be recognised manually.

Collocations three words to the left and right of each item are calculated by *Microconcord* (Johns - Scott 1993) and then a mutual information score based on the logarithm of the observed frequency of collocation divided by the expected frequency of collocation is calculated. This gives a list of collocates as in Appendix 4, the most frequent collocates of the item *of* from the abstracts subcorpus. This process is repeated for each item, and the results are discussed below. Readers will note from Appendix 4 that *of* collocates significantly with itself, because there may be up to two intervening words before the next occurrence of *of*. Readers are invited to read the instructions for each Appendix after Appendix 1, noting also that the symbol _ before an item indicates that it is a right-hand collocate.

4. . Transitivity processes and phraseology.

One major result emerges from the data and needs to be signalled here. There is a strong tendency for phraseology to be structured by lexical items that share semantic characteristics. We have already mentioned these terms before, but we can now summarise four process types that correspond in nature (but not in kind) to Halliday's processes of transitivity:

a) research processes (cognitive, verbal processes) characterise the writing activity or act of observation that the researchers are engaged in (From the Medline titles corpus: study, evaluation, case, comparison, analysis, detection, characterisation, assessment).
b) clinical processes (material) include the medical or methodological processes which subjects (patients, mice etc.) receive: (From the Medline titles corpus: treatment, therapy, care, management, resection, injection).
c) empirical processes (relational, material) characterise theoretical models or chemical interactions (From the Medline titles corpus: effect, role, risk, stability, influence, use, relevance, increase).
d) biochemical processes (material) label the interaction of biochemical entities: (From the Medline titles corpus: expression, infusion, synthesis, hydrolysis, induction).

So called 'regular' phraseological units typically restrict the semantic components of the phrase to one process type (or even one subtype). This is in effect the principle behind the Cobuild dictionary: senses are defined by phraseology. We use this classification to describe the global characteristics of a phrase. But we emphasise here that these categories emerge from the corpus analysis and therefore need to be considered in their phraseological environment since one of the defining characteristics of each process type is that they occur in complementary distribution to each other.

5 Data : Phraseology in cancer research abstracts

There are 29 136 words in the PSC abstracts subcorpus. Wordlist data reveal the following salient items:

Table 5. *Wordlist* salient items in the PSC abstracts subcorpus.

Frequent salient items (>500 tokens) Salient grammatical items

1	tumor (114/1235)	but (67/663)
2	expression (63/584)	these (119/1399)
3	patients (63/582)	of (1367/21 309)
4	induced (57/521)	there (40/ 444)
5	but (67/663)	in (912/ 14 349)
6	growth (69/707)	was (304/ 6 146)
7	cancer (54/522)	that (227/3357)
8	these (119/1399)	did (34/395)
9	tumors (82/903)	who (14/129)
10	treatment (59/606)	both (55/713)

Abstract-salient lexical items are largely disease-related entities (*mammary, tumor*) or cellular processes (*expression, induced*). In particular, important processes involving tumor growth appear to be the most frequent items in the abstract (*heterozygocity, growth, expression, active, cancer*). Not represented in the top ten, but equally relevant from the first 100 significant lexical words are items indicating a general description of the shape of the data rather than the methods (*correlated, decreased, increased, interval*, *level*) and verbs that report past research (*studied, suggest*) and this tendency is borne out by the phraseology.

5.1 Abstract salient item 1: But

The very high significance of *but* (compared with other grammatical items in abstracts) suggests that the reporting of negative results is a fundamental characteristic of abstracts. In particular 'but' is an explicit signal of reversal and evaluation of the direction of quantifiable results (up, down or stable):

but displayed no significant reduction...
but this also fell...
but decreased sharply...
but restabilized...
but adjusted to milder in vitro expression...

Subjects of clauses introduced by *but* are all related to the measurement of the efficiency of drugs (*patients* is a frequent left-collocate, and other items include *resistance, efficacy, immune response*). In results sections on the other hand, we find that the tendency is to explain negative results or to state negative empirical processes rather than quantify them (*however...did not correspond, although this did not result in...*). To summarise, in abstracts negative data is quantified whereas in results sections negative data can be seen to be 'qualified'.

5.2 Abstract salient item 2:these

'This' functions to signal refocussing and rephrasing reformulation. This function is shared by Discussion sections and a more detailed analysis is seen out in our discussion of 'this' (Gledhill 1995b). We note here that 'these' differs from 'this' (in discussion sections) in that almost half of the occurrences of *these* are as pronouns introduced by *of*, while 'this' is mostly a modifier. The referents of *these* tend to be very specific illness-related items (*carcinogenic factors, leucocytes, oncogenes, metastases*) and items that introduce *of* are items of measurement (*half of these, the majority of these, concentrations of these*) a pattern that coincides exactly with the one set out below for *of*. This indicates a correlation with our earlier finding that abstracts tend to favour the use of deictic refocussing encapsulation. The high significance of *these* (according to Appendix 2) here also coincides with Nwogu and Bloor's (1991) observation that abstracts tend to employ simple thematic progression, linearly converting rheme to theme.

5.3 Abstract salient item 3:of

In the subcorpus of titles (Gledhill 1995b), *of* was seen to play a key role in nominal groups with a typical treatment-*of*-disease pattern. Such a symmetrical solution-problem pattern is expanded in the abstract, the major difference being that while items in the title corpus tend to predict *of* with no strong right-collocates, in the abstract there are just as many significant right-collocates, such as *human, these, was*. Another difference with Titles is that Abstracts involve the quantification or description of disease, where *o f* introduces the semantic 'support' (not necessarily 'head'): *number, concentration, levels, incidence, frequency, majority*, *presence ... of... cancer, tumour, oncogene, growth, expression, patients, mice, human*. A second pattern tends to introduce empirical items that explain the potential treatment of the disease (*effect, role, mechanism, treatment, inhibition, synthesis... of.. drug X, doxorubicin, compounds*). As the first element becomes more necessary to the interpretation of the next item, the phrase introduced by *of* in the second group can be seen, in Sinclair's terms (1991:82-83) as 'focus' rather than 'support'. These patterns can be seen in example concordances in the Appendix.

The 'treatment-*of*-disease' pattern can be seen as an overriding pattern, but within this there is considerable phraseological change. We have identified four different problem-solution patterns of complex stereotypical phraseology with *of* for some of the most frequent left-

collocates of *of* in the Abstract: (*effect, loss, number, presence*) and there does not seem to be any evidence to suggest that any such middle frequency item (often termed sub-technical items: Francis 1993) shares the same phraseology as any other. In particular, the solution-problem / treatment- disease pattern seen in the title does not appear to be fixed for each item in the abstract. For example, *presence of* has an alternative pattern if post-modified: *the role/ presence of* (*drug X*) *in* (*illness Y*). Other items require more explicit modification. *Effects* and *effect* are usually in subject position and are almost always pre-modified by a treatment-oriented item (*growth-inhibitory, antitumour, chemopreventive*, *protective*) or a research-observation item indicating some problem (*adverse, side-effect, toxic*). On the other hand, *presence* is often used in a prepositional phrase functioning as adjunct, (preceded by *in, for, on*) or in a subordinate clause where there is no explicit statement of problem or solution, and where *presence of* signals an illness-related specific item where a possible link with cancer is being explored: *retrovirus, ras proto-oncogenes, maternal toxicity*.

In addition, the expression *use of* represents one of the most stereotypical patterns of the abstract. It is always preceded by some degree of measure or a methods-oriented specification of use (*daily, widespread, regular, intensive, combined, clinical, potential*) and followed by a specific drug X(1) and an expansion of the treatment and illness (*with drug X(2), in the study of illness Y*, *in the treatment of, in the evaluation of Y*) and finally followed by some degree of evaluation or a research process: *resulted in..., should be considered, is discouraged, is discussed*.

In a different kind of distribution, the significant collocate *loss* appears to have become terminologised in the fixed expression *loss of heterozygocity*. *Loss* also appears in thematic position where a research statement is phrased in the passive or placed after the term (*loss of X...was found, occurred*, *occurring*), although there are reporting instances such as *suggest that* .... which contradict this. The pattern occurs more regularly with *effect/s* where specific reporting items are sometimes placed as hedges: (*effect/s of X... were found, reduced, appeared to be.., as shown..., and seem to...*). Interestingly, among most of the measurement-illness phrases mentioned above, the reporting verb precedes the expression (*shows/ confirms/ indicates ...the presence of*, *incidence of, absence of*). A fourth pattern is represented by the expression *number of* which is not immediately preceded or followed by a reporting discourse item. It may be that there is a differentiated pattern of phraseology in which *of* has a role as constructor of nominalisations of measurement and qualification (i.e. the first use mentioned above), in conjunction with expressions of research reporting and

evaluation (the second use). The writer can thus choose to emphasise the 'self evidence' of the data by evoking phrases involving *number of*, or may wish to thematicise the study and be required to use stereotypical measurement-disease phrases, or alternatively thematicise the results and use an expression with items such as *effects*.

5.4 Abstract salient item 4:there.

'There' reveals a prevalence for existential process clauses in the Abstract, most often expressing explicit evaluation of the shape of research articles' results (up, down or no change). In the Abstracts subcorpus, the dummy pronoun *there* is uniquely followed by *was* and *were* and occurs in thematic position after a statement of methodology. The (quantitative) empirical concern for the overall direction of the data in the abstract is invariably pre-modified by explicit evaluation:

Existential process:	*Evaluated quantification*:
there was/ were...	no difference, no significant difference, a reduction in the percentage of, considerable variation, a transiently increased number of correlations, strong correlation, no change, pronounced distribution decreased hepatocyte labelling, a high degree of similarity

These expressions typically precede the highly significant items within the subcorpus that deal with statistical direction or relation (as indicated by the right-collocates of *there*: *increased, decreased, interval, correlated*). There are one or two exceptions to the pattern, where empirical items are qualitative rather than quantitative, for example:

there were/ was...	pronounced effects no complete response clearly a strong genetic predisposition...

5.5 Abstract salient item 5:in

'In' is used most frequently in three patterns:

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 relational verbs (*accumulates in, is low in, resistance was narrower in the cell*), or as a phrasal element in research processes (*observed, detected*)
3) to introduce 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 other sections, for example in titles, the most frequent use of 'in' is its spatial meaning (*in the liver, in cells*) (Gledhill 1995b). 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' items: cancers, subjects or specific disease-related entities (*cancer, breast cancer, tumor-bearing animals, patients, tumor-bearing mice, cytokines, methylene chloride*). It is likely that reference and other discoursal factors have a role to play in this distinction. But both these uses are of generic *the* in prepositional phrases and Master (1987) has claimed that discoursal factors (while crucial elsewhere) do not affect generic article / zero-article usage. So an alternative explanation may be that just as article usage is highly idiomatic in certain specific semantic domains in the general language, then it may be that phraseology becomes more idiomatic in the specific language.

5.6 Abstract salient item 6: was

The simple past is the preferred tense for presenting the research article's present methodology and results. Ironically, as we have seen, the present is used to introduce previous research. This is contrary to existing statements on the subject (Hanania and Akhtar 19985) and to Malcolm's (1987) distinction (past for generalisations, present for specific data). 'Was' reports the research article's (clinical) methodology and non-quantitative (empirical process) results. 'Was' in the abstract can be seen to play a completely different role to its present tense version: *is*. In the abstract, there are two patterns for *is*:

1) *There is*... followed by a statement of evidence: *no evidence, no molecular evidence, no indication, no significant difference +that, for this, to suggest etc,*
2) Extraposed *it* and a *that*-clause: *it is ...concluded, apparent, desirable, essential, important, possible, believed, expected, likely that*…followed by a statement of findings.

Was does not share any of these phraseological characteristics, and is instead involved with statements of qualitative results where the subjects are either key biochemical entities in the cell (*peripherin, protein, nucleus, DNA, glycoprotein, toxicity*) or biochemical items involved with tumour's effect on the metabolism (*growth, weight, vasodilation, expression*). As in Methods sections, *was* introduces some passives with technical verbs as past participles which are often pre-modified by a technical adverb:

was	*metabolically expressed*
	immunologically reacted
	enzymatically deaminated
	induced
	carried

However, the majority of passives in the abstract are more empirically or research process oriented and resemble passives in results sections:

was (research process):
.... observed, found, detected, determined, studied, seen, shown investigated, demonstrated, performed, established, confirmed, compared.

5.7 Abstract salient item 7:that

Across the PSC corpus, 'that' as complement is used to play an important role in reformulating the claim as a cognitive research process (*The idea that, we conclude that*). A frequent use of 'that' in abstracts is in extraposed *it* clauses following verbs of cognition and belief (*it is ...believed, expected, concluded ... that*) or adjectives of possibility or volition (*important, possible, likely, desirable, evident*). Similarly reporting clauses have clear limitations on the subject of the clause:

we conclude that
we find that

while more data-oriented items introduce *indicat*e,

values indicate that

findings indicated that
results
information

while *studies* and *results* also introduce *demonstrated*. A similar pattern is observed in discussion sections. One difference with the discussion section is the important amount of 'that' functioning as relative pronoun in embedded clauses. It functions by referring most often back to a specific chemical and establish some characteristic function of the entity: (Z occurred to *chemical X* that is...*normally responsible for, typical, expressed only as, effective in maintaining levels of*) or emphasise the status of the knowledge structure (*allow prediction of experimental factors that underline our lack of understanding of these processes*).

5.8 Abstract salient item 8: did

Did is only used in two ways in the corpus: to introduce the negative, *not*, and in elliptical expressions such as *as did the*... Perhaps surprisingly, the presentation of negative results is a key function in Abstracts and we assume that they are emphasised (as we have seen for *but*) partly to deflect possible criticism but also because empirical negative results are just as newsworthy in the demolition of null-hypotheses.

The subjects of *did* reflect the typical sentence themes of the abstract: processes of tumour growth (or stopping the growth) (*propagation, growth, expression, inhibition*) and pharmaceutical molecules that are involved in helping or hindering these processes (*cholesterol, methyl chloride, doxorubicin, heparin*). Verbs that are negated tend to be the measurement or reporting verbs prevalent after 'but' in the abstract (did not… *increase, decrease, show that*). Typical subjects of these clauses are biochemical processes (*efficiency, correlation, the data, sample response*). Again, this pattern is not reflected in results sections where negative results relate to empirical processes of causality rather than quantification. There is little evidence to suggest that researchers want to 'hide' negative evidence: negative results in themselves are not necessarily *bad*, they may well support the writers' research hypothesis. The reason for the difference in expression may be that results sections need to explain negative process results (such as lack of causality, effect or evidence) while abstracts state data-related results, leaving inferences about 'higher' empirical or research implications to the reader.

5.9 Abstract salient item 9: who

Who refers to the only participants other than the researchers (*we*) who appear in the corpus: the *patients* and analogous terms such as *physiological group, those*... Consequently, relative clauses introduced by *who* deal with the role of *patients* as subjects (in the grammatical and clinical sense) who are seen as active recipients of research, rather than objects to be experimented on:

subjects who receive active management
patients who had received active management
% of those who had taken aspirin,
subjects who took part in radiation studies
patients who showed positive response to the administration of AZT
those who progressed slowly
cancer patients who succumbed
patients who had tumours,

In particular, patients are never *given* drugs, they receive them (*who receive carboplasmin, receive Doxo, receive doxorubicin*). This is quite a clear example of the way phraseology helps to shape a specific view of a transitivity at the same time as framing terms stereotypically. For example, given that all object complements of the verb *'receive'* are drug treatments, the non-initiate observer is compelled to assign a similar semantic profile to the terms *receive active physiological management* and *receive administration*. The phraseology of the term *management* (the 46th most frequent term in the PSC corpus) allows us to establish its meaning within the corpus not as what one might expect ('organisation of personnel') but as part of a recurrent transitive structure involving patients and 'receiving' - the preferred phraseology for the experimental application of drugs *in vivo*. While 'take part in' and 'receive' are the most common formulations after 'who', the same phraseology is not reserved for the other participants in the process. Animals tend to be 'given' drugs, so we find (especially in the methods section) 'mice were exposed to, fed, given...'. We did find, however, one instance of mice infelicitously 'taking part' in an experiment:

mice who took part in the control study were given doxorubicin based analogues.

5.10 Abstract salient item 10: both

In many of the cases where 'both' is used as a linking conjunction, it is largely redundant. The following sentence is typical:

Two antibodies that inhibited both anchorage dependent and anchorage independent growth also blocked...

One explanation may be that 'both' is considered necessary by the researcher to emphasise two complementary alternatives, thus establishing a basic taxonomy. In abstracts we find the following oppositions:

both accelerate and delay,

pre-B	early cells
high	low secretors
mouse	human
rats	mice
cytosolic	particulate functions
oxidative	reductive metabolism
destructive	regenerative processes
normal	tumor cells

This set of oppositions, left implicit by the writers, provides us with a set of fundamental oppositions that allow us to situate them in relation to other other concepts and terms in the corpus and to further define the discipline.

6 Discussion: Phraseology and discourse.

On the basis of the phraseological analysis set out above, we claim that collocational patterns represent an implicit model of the most common phraseological choices available to the cancer researcher in the specific subgenre of the abstract, and these choices are limited by the topic and some sense of the preferred direction that the phrases may take as a longer string. We shall refer to the system of choice where a momentum of direction is maintained throughout the system and where collocations join each other as *collocational cascades*. What is significant about this is that the cascades represent the 'generic' part of the abstract for the reader: elaboration, specification, reformulation and other deviations from the cascade will undoubtedly attract the attention of the reader and ultimately determine what changes the research paradigm.

Another important finding is the semantic corresponence between collocation and As Francis says:

> "As we build up and refine the semantic sets associated with a structure, we move closer to a position where we can compute a grammar of the typical meanings that human communication encodes, and recognise the untypical and hence foregrounded meanings as we come across them." (Francis 1993:155).

The patterns we have identified in the analysis of the abstracts subcorpus are not accidental. There is now a body of linguistic theory that sees such patterns as central to the way discourse is *construed*, or to reformulate Halliday (1995), how we build and interpret the world through discourse. This view of language sees the semantics

of the word as textually distributed and syntax as intimately linked with lexical knowledge. Fillmore, Kay and Connor (1988). write of phraseology in terms of:

> ...phenomena larger than words, which are like words in that they have to be learned separately as individual facts about pieces of the language, but which also have grammatical structure [and] interact in important ways with the rest of the language. (1988:504)

In the specific context of cancer research abstracts, such instantial knowledge involves how to introduce the disease and its treatment as well as stereotypical patterns such as the use of *use*, *active physiological management,* and even a subconscious knowledge of duality in the discipline introduced by *both*. Although this is only a sample, these patterns can be seen to be important processes of writing and reading abstracts in this specialist field. In this regard, Francis (1993) has argued that such lexical knowledge is a key mechanism by which we progress from ideas to linguistic form:

> As communicators we do not proceed by selecting syntactic structures and independently choosing lexis to slot into them. Instead we have concepts to convey and communicative choices to make which require central lexical items, and these choices find themselves syntactic structures in which they can be said comfortably and grammatically (1993:122)

Given this view, that meanings acquire their won wordings, we can therefore conceive of phraseology as linguistic forms motivated by rhetorical aims and which further shape the text that is to follow. The next logical step it to analyse grammatical metaphor and the way it shapes concepts throughout a running text, and this is under way (Gledhill, forthcoming).

What are the benefits of recurrent phraseological patterns in the business of scientific writing? We have hypothesised that phraseology is a key process that corresponds to conventional writing strategies in the research articles we have been studying. While rhetorical structure allows for accurate prediction on a broad scale, phraseological patterns could also be involved in allowing for browsing and skimming though a text, as Nystrand suggests (1986). Previous research has addressed the question of non-linearity (the in-built mechanisms of the text to allow for skimming and partial reading) by invoking a notion of 'rhetorical convention'. In their studies of signalling and use of rhetorical structure, Swales (1981), Nwogu (1989) and Sharp (1989) had found that predictable elements of rhetorical structure and visual format help readers

to identify where to jump to, to guess the content of conventional areas of the texts. But while such analysis helps describe what one might call the A-B-C reading of texts, it doesn't account for how scientists can make a coherent account of a partially read text, or how parts of the text may be considered to be cohesive even at some distance apart, a notion that Hoey (1991) has been exploring in terms of relating sentences with similar lexis. So in addition to powerful tools of rhetorical structure and format, it is worth considering, for example, whether grammatical parallelism, conventionalised phrases and cohesive networks might also be used in these texts to complement their non-linearity of use.

It is also our hypothesis that these phraseological patterns are acquired piecemeal by the slow processes of re-editing and rereading that apprenticeship in the discourse community requires (Myers 1990). The whole pghrase acquisito of language is certainly a process that has been proposed for the general language by such researchers as Pawley and Syder (1983), Peters (1983) and Widdowson (1989). It may be possible, for example, for genre analysis, or ESP to make some use 'collocational cascades' as short cuts in order to save time when teaching English for reading and producing research articles as Johns (1993) has suggested. In addition, the slower, immersed approach to acquiring phraseology may be a useful analytical tool, not only in monitoring the linguistic progress of apprentice writers, but also in analysing how texts are edited, how coherence develops chronologically throughout a text and how phraseology evolves over time, just as Atkinson has demonstrated with rhetorical structure (1992). It is possible that such a research paradigm already exists, and although we term it 'developmental linguistics' here, how its potential is to be realised is another story.

References. AARTS J. and MEIJS W. (eds.) 1986 Corpus Linguistics II Amsterdam: Rodopi

AHMAD K., FULFORD H., GRIFFIN S. and HOLMES HIGGIN P. 1991 "Text-based knowledge acquisition- A language for specific purposes perspective." Guildford: ESPRIT II Report for the University of Surrey.

AIJMER K. and ALTENBERG B. (eds.) 1991 English Corpus Linguistics London: Longman

ATKINS S., CALZOLARI N. and PICCHI E. 1992 "Computational lexicography." Pre-Eurolex Tutorial University of Tampere, Finland, August 4-9, 1992

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

ATKINSON D. and BIBER D. 1994 "Register: A review of empirical research." in D. Biber and E. Finegan (eds.) 1991 :1-68

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

BAKER M., FRANCIS G. and TOGNINI-BONELLI E. (eds.) 1993 Text and Technology Amsterdam: John Benjamins

BERNIER C.L. 1985 "Abstracts and Abstracting." in DYM :423-444

BIBER D. 1988 Variation across Speech and Writing Cambridge: Cambridge University Press

BIBER D. 1992 "Using computer-based text corpora to analyze the referenial strategies of spoken and written texts." in J. Svartvik 1992 215-252

BIBER D. and FINEGAN E. 1986 "An initial typology of English text types." in J. Aarts and W. Meijs 1986 :19-46

BIBER D. and FINEGAN E. (eds.) 1991 Sociolinguistic Perspectives on Register Oxford: Oxford University Press

BRETT P. 1994 "A genre analysis of the results sections of sociology articles." In English for Specific Purposes journal Vol.13/1 pp47-59

BUTLER C. 1985 Statistics in Linguistics Oxford: Basil Blackwell

CLEVELAND D.B. and CLEVELAND A.D. 1983 Introduction to Indexing and Abstracting Princeton Colorado Libraries Unlimited

CREMMINS E.T. 1982 The Art of Abstracting Philadelphia ISI Press

DIODATO V. 1982 "The occurrence of title words in parts of research papers:

variations among disciplines." in Journal of Documentation Vol. 38/3 :192-206

DRURY H. 1991 "The use of systemic linguistics to describe student summaries at university level." in E.Ventola (ed.) 1991: 431-456

ENDRES-NIIGGEMEYER B. 1985 “Referierregln und Referate- Abstracting als regelsgesteuerter Textverarbeitungsprozeβ.” in Nactrichten für Dokumentaristen Vol. 36/1 :38-50

ENDRES-NIGGEMEYER B. 1990 "A procedural model of an abstractor at work." in International Forum of Information and Documentation 15/4: 3-15

ENDRES-NIGGEMEYER B., WAUMANS W. and YAMASHITA 1991 "Protocol analysis of non-native abstractors." in Text Vol. 11/4 :523-552

FILLMORE C.J., KAY P. and O'CONNOR M.C. 1988 "Regularity and idiomacy in grammatical constructions." in Language Vol. 64 :501-538

FLØTTUM K. 1985 "Methodological problems in the analysis of student summaries" in Text Vol. 5/4 :291-308

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

GIBSON T.R. 1992 "Towards a discourse theory of abstracts and abstracting." Unpublished Ph.D. Thesis, English Language Department, Nottingham

GLÄSER R. 1991 "The LSP genre abstract - revisited." in ALSED - Newsletter Vol. 13/4 :?

GLEDHILL 1994 "La Phraséologie et l'analyse des genres. L'exemple des formules rhétoriques dans *Le Mond*e" Papers of the Institute for the Study of Discourse in Society, Department of Languages and European Studies, Aston University.

GLEDHILL 1995 "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 1995 Scientific innovation and the phraseology of rhetoric. Posture, reformulation and collocation in cancer research articles. PhD. Thesis, Department of Languages and European Studies, Aston University.

GOPNIK M. 1972 Linguistic Structures in Scientific Text Den Haag: Mouton

GUBA E.G. and LINCOLN Y.S. 1982 "Epistemological and methological bases of naturalistic inquiry." in Educational Communication and Technology Journal Vol. 30/4: 233-252

GUNAWARDENA C.N. 1989 "The present perfect in the rhetorical divisions of biology and biochemistry journal articles." in English for Specific Purposes journal Vol. 8/3 :265-273

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

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

HANANIA E.A.S. and AKHTAR K. 1985 "Verb form and rhetorical function in science writing: a study of MSc theses in Biology, Chemistry, and Physics." in English for Specific Purposes journal Vol. 4 :49-58

HOEY M. 1991 Patterns of Lexis in Text Oxford University Press

HOWARTH P. 1993 "A phraseological approach to academic writing," in Review of English Language Teaching Vol. 3/1 1993: 58-69

HUTCHINS J. 1977 "On the structure of scientific texts." University of East Anglia Journal of Linguistics No. 5

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 22nd 1993

JOHNS T. and SCOTT M. 1993 Microconcord: Concordancing Program, Oxford University Press.

JOURNAL OF THE CHEMICAL SOCIETY (JOC) PERKIN TRANSACTIONS 1993 "Instructions for Authors" in Journal of the Chemical Society (PRB4) Vol.1/164 vii-xxviii Washington NY: The American Chemical Society

KNORR-CETINA K.D. 1983 (ed.) Science observed : perspectives on the social study of science London : Sage

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

LANE P. 1992 La Périphérie du Texte. Nathan: Paris.

LOVE A. 1992 "Lexico-grammatical features of geology textbooks " in English for Specific Purposes journal Vol.12/3: 197-218c

MAIZELL R.E., SMITH J.F., and SINGER T.E.R. 1971 Abstracting Scientific and Technical Literature London: Wiley Interscience

MALCOLM L. 1987 "What rules govern tense usage in scientific articles?" in English for Specific Purposes journal Vol. 6/1 :31-43

METANOMSKI, D. 1993 (Editor in Chief- Chemical Abstracts Service) *Personal communication*.

MASTER P. 1987 "Generic *the* in *Scientific American*" in English for Specific Purposes journaVol. 6/3 :165-186

MEYER P.G. 1988 "Statistical text analysis of abstracts: A pilot study on cohesion and schematicity." in Computer Corpora des Englishen Vol. 3 :17-40

MOON R. 1992 "The is a 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 University of Wisconsin Press

NWOGU K.N. 1989 Discourse variation in medical texts: Schema, theme and cohesion in professional and journalistic accounts." Unpublished PhD. thesis, Language Studies Unit, Aston University.

NWOGU K. N. and BLOOR T. 1991 "Thematic progression in professional and popular medical texts." in E. Ventola (ed.) 1991 :369-384

NYSTRAND M. 1986 The Structure of Written Communication: Studies in Reciprocity between Writers and Readers Orlando Fl.: Academic Press

OSTER S. 1981 "The use of tenses in reporting past literature in EST." in English for Academic and Technical Purposes: Studies in Honour of Louis Trimble L. Selinker, E. Tarone and V. Hanzeli (eds.), Massachussets: Newbury House :76-90

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

RENOUF A. and SINCLAIR J. McH. 1991 "Collocational frameworks in English." in K. Aijmer and B. Altenberg (eds.) 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 Nardstetter Verlag

SALAGER-MEYER F. 1992 "A text-type and move analysis study of verb tense and modality distribution in medical English abstracts." in English for Specific Purposes journal Vol. 11/2 :93-114

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

SASTRI M. 1968 "Prepositions in chemical abstracts." in Linguistics Vol. 38 :???

SAVILLE-TROIKE M. 1982 The Ethnography of Communication Oxford: Basil Blackwell

SCIENCE CITATION INDEX 1993 Permuterm Subject Index

SCOTT M. 1993 ""Lexical tools for genre analysis by computer." paper presented at the BAAL annual meeting, Salford University 14-16 Sept. 1993

SELINKER L., TARONE R. and HANZELI V. (eds.) 1981 English for Academic and

Technical Purposes: Studies in Honor of Louis Trimble Newbury House: Mass. USA

SHERRARD C. 1989 "Teaching students to summarize.: Applying textlinguistics." in System Vol. 17/1: 1-11

SINCLAIR J. McH. (ed.) 1987 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. 1993 "Text corpora: Lexicographer's needs." in Zeitschrift für Anglistik und Amerikanistik Vol. XLI: 1/1: 5-13

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

SVARTVIK J. (ed.) 1992 Directions in Corpus Linguistics Proceedings of the Nobel Symposium 82: Stockholm 4-8 August 1991.

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

TADROS A. 1985 Prediction in Text Discourse analysis monograph No. 10, English Language Research, University of Birmingham

VENTOLA E. (ed.) 1991 Functional and Sysemic Linguistics: Approaches and Uses Den Haag: Mouton de Gruyter

ZAMBRANO S. 1987 "A Comparison of the Linguistic Features and Discourse Structure of Abstracts and Conclusions" unpublished MSc Thesis, Language Studies Unit, Aston University

Appendix 1: The Topical breakdown of the PSC Corpus.

Oncology *(Cancer Research* Total=83 articles)

Topic	No. of Articles	Explanation.
Chemotherapy:	26	Chemico-toxic effects on cancer.
Carcinogenesis:	18	Processes that activate cancer.
Histopathology:	12	Metabolic effects of tumours.
Immunohistochemistry:	11	Organic resistance to tumours.
Cytogenetics:	10	Genetic characteristics of cancer.
Cancer Epidemiology:	2	Population study of carcinogenesis.
Radioimmunology:	2	Radio-toxic effects on tumours.
Histology:	1	Organic properties of tumours.
Immunology:	1	Organic resistance to tumours.

Pharmaceutical science *(Medicinal Chemistry* Total=*63)*

Structural chemistry:	18	Processes of chemical interaction.
Organic Chemistry:	15	Functions of organic compounds.
Toxicology:	13	Effects of drugs on metabolism.
Pharmacology:	9	Effect of drugs on disease.
Enzymology:	8	Organic compounds in the metabolism.

General Medicine (Total=4)

Epidemiology:	1	Population study of disease.
Gynaecology:	1	Population study of fertility.
Patient Care:	1	Hospital management of disease.
Virology:	1	Population study of rubella virus.

Instructions for Appendices 2, 3 and 4
APPENDIXES **2-3**

These are extremities of one raw *Wordlist* output file (Mike Scott 1994), Appendix 2 indicating the first 40 items in the list (typical of the abstract) and Appendix 3 the last 40 (least typical of the abstract and therefore associated with the rest of the research article).

First Column: RANK = Not the rank of the word in the corpus: but its rank as a word that is more significantly frequent in the abstract than in the rest of the article.
3rd/4th Column: PSC Abstracts Freq. = The frequency of the word in the abstracts (with a percentage of the whole corpus abstract if over 0.1 %).
5th/6th Column: PSC Main corpus Freq. = The frequency of the word in the corpus as a whole (with a percentage of the whole corpus abstract if over 0.1 %).
7th Column: X2 = (Chi-squared) Significance score based on the following equation:

$$\chi^2 = \frac{\text{(the sum of) } (O-E)^2}{E}$$

where O = observed word frequency of a word in the abstract..
and E = expected word frequency, based on the number of words in the abstract, multiplied by the frequency of the word in question divided by the total number of words in the entire corpus (500 000).
8th Column: Significance p = A significance score at 1 degree of difference (k-1) based on the Chi-score where p = 0.05 would be 'significant at the 5% level i.e. very highly significant. As can be seen in Appendix 2 , items such as *suggest* and *but* are extremely significant with a significance at less than the 0.1% level. Where there are not enough words to justify a significance score no *p=* score is indicated, although the *X2* score may place these items very highly (for example: the word *abstract* itself with only 32 instances in the corpus as a whole is judged to be highly significant because it occurs 32 times, of course, in the abstract).

APPENDIX 4

In this table, each collocate of the word *of* is listed on the left, the most significant collocate at the top, the least at the bottom. The calculations represent the Mutual Information of collocation, introduced by Atkins et al. (1992). The sum of probabilities (P (x+y)) of two items are divided by their chance of collocating at random (P (x*y)), this is compared on a logarithmic scale to base 2.

Appendix 2:* Wordlist *(Scott 1993) list of the 40 Most Frequent 'Abstract-Keywords' from the PSC corpus.
That is: Words which are most typical of the abstract.

		PSC abstracts.		PSC main corpus.		Significance.	
RANK	WORD	Freq.	%	Freq.	%	X2	p =
1	ABSTRACT	32	(0.1%)	32	234.6		
2	SUMMARY	39	(0.1%)	63	203.3	0.000	
3	DOXORUBICIN	26		97	54.7		0.000
4	5FU	14		45	34.1		
5	MYOD1	9		19	33.2		
6	DOXO	16		59	33.0		
7	KG	43	(0.1%)	303	30.4	0.000	
8	SUGGEST	30	(0.1%)	177	30.3	0.000	
9	HN9	5		5	29.9		
10	H691VDS	5		6	26.4		
11	HETEROZYGOSITY	13		50	24.8		
12	ESTERS	12		44	24.2		
13	MAMMARY	26		161	23.7	0.000	
14	ACTIVE	33	(0.1%)	231	23.4		0.000
15	DOSES	29		193	22.8		0.000
16	STUDIED	26		164	22.8		0.000
17	RESISTANEE	4		4	22.4		
18	SPIRAMYEIN	4		4	22.4		
19	TUMOR	114	(0.4%)	1235	21.8		0.000
20	INHIBITED	21		121	21.7		0.000
21	IOA	6		12	21.7		
22	EXPRESSION	63	(0.2%)	582	21.6	0.000	
23	PATIENTS	63	(0.2%)	584	21.3		0.000
24	CORRELATED	13		56	21.0		
25	MHB	16		80	20.8		0.000
26	ACYLOXYBENZYL	9		29	20.7		
27	ANTHRACENE	13		57	20.5		
28	INDUCED	57	(0.2%)	52	20.1		0.000
29	OA	4		5	19.2		
30	NDENT	5		9	19.0		
31	BUT	57	(0.2%)	663	18.1		0.000
32	IMMORTALIZED	13		69	17.9		
33	SHOWED	43	(0.1%)	375	17.4	0.000	
34	INCREASED	43	(0.1%)	376	17.2	0.000	
35	INTERVAL	12		56	16.9		
36	PDL	4		6	16.7		
37	GROWTH	69	(0.2%)	707	16.4		0.000
38	DECREASED	23		161	15.9	0.000	
39	CANCER	54	(0.2%)	522	15.7		0.000

Appendix 3:* Wordlist *(Scott 1993) list of the 40 Least Frequent 'Abstract-Keywords' from the PSC corpus.
That is: words which are more typical of the article.

5061	DBA	1		115		5.8	0.016
5062	TEMPERATURE	4		204		5.9	0.015
5063	WOULD	5		232		6.0	0.015
5064	NM	4		206		6.0	0.015
5065	X	42	(0.1%)	1045	(0.2%)	6.0	0.014
5066	F	4		210		6.2	0.013
5067	NMR	2		158		6.4	0.011
5068	MIXTURE	3		188		6.5	0.011

5069	G		33 (0.1%)	878 (0.2%)	6.6	0.010
5070	GEL		1	143	7.4	0.007
5071	BECAUSE	3	205	7.4	0.006	
5072	TEST		5	262	7.5	0.006
5073	D		29	821 (0.2%)	7.6	0.006
5074	STANDARD	2	182	7.8	0.005	
5075	IS		146 (0.5%)	3169 (0.6%)	8.1	0.004
5076	J		3	223	8.4	0.004
5077	REPORTED	9	395	9.0	0.003	
5078	CONTAINING		8	370	9.1	0.003
5079	I	85 (0.3%)	2029 (0.4%)	9.4	0.002	
5080	THE		1574 (5.4%)	29122 (5.8%)	9.5	0.002
5081	WASHED	1	190	10.1	0.001	
5082	BUFFER		5	313	10.3	0.001
5083	OBTAINED	18	640 (0.1%)	10.3	0.001	
5084	EACH		16	595 (0.1%)	10.4	0.001
5085	O		8	397	10.4	0.001
5086	M		32 (0.1%)	973 (0.2%)	11.0	0.001
5087	DESCRIBED	9	436	11.1	0.001	
5088	FOR		246 (0.8%)	224 (1.0%)	11.2	0.001
5089	SHOWN		21	731 (0.1%)	11.2	0.001
5090	CM		5	345	12.0	0.001
5091	MEDIUM		6	376	12.2	0.000
5092	INCUBATED	1	237	12.9	0.000	
5093	MIN		19	725 (0.1%)	13.1	0.000
5094	H		75 (0.3%)1	961 (0.4%)	13.5	0.000
5095	MMOL		2	302	14.7	0.000
5096	SOLUTION	6	428	15.0	0.000	
5097	ADDED		3	340	15.1	0.000
5098	IT		29	1006 (0.2%)	15.2	0.000
5099	HZ		1	294	16.2	0.000
5100	MM		9	540 (0.1%)	16.6	0.000
5101	THEN		4	420	17.9	0.000

Appendix 5: Sample ordered concordance of the word 'of' from the PSC abstracts subcorpus.

1 antitumor drug. Specific side effects	of	Doxo primarily affect the cardiac m
1 rtly responsible for the toxic effect	of	Doxo on cardiac muscle and that loc
1 analysis. The growth-inhibitory effect	of	doxorubicin, daunorubicin, N,N- dim
1 ely inhibited the permeability effect	of	endothelin-l in the stomach and duod
1 TA) receptors in mediating the effects	of	endothelin-1 on microvascular perme
1 was obtained. The anti- tumour effects	of	gemcitabine appeared to be similar o
1 wn the cancer chemopreventive effects	of	green tea in several animal tumor m
1 by DEN and BP, the protective effects	of	GTP were between 38-43 and 25-46% r
1 e, 2-5 months). The main side effects	of	ifosfamide were alopecia (83% of pa
1 uld be continued to monitor the effect	of	introducing the measles, mumps, and
1 starch. There was a pronounced effect	of	molecular weight of PAA on the bioa
1 tration added to the antiemetic effect	of	ondansetron principally in patients
1 ody-specific targeted radiation effect	of	RIT was seen. '3'1-labeled MB-I pro
1 pendent manner (IC50 140 nM). Effects	of	several antiarrhythmic drugs on Y-26
1 in-induced nephrotoxicity. The effect	of	the simultaneous administration of
1 intaining the cardioprotective effect	of	the liposome carrier as suggested b
1 ed with the growth inhibitory effects	of	the different gents and seem to ref
1 could arise from the metabolic effect	of	the tumour on host tissues, mediated
1 icantly reduced the therapeutic effect	of	the treatment. The clinical relevan
1 ed the maturationpotentiating effects	of	the bile acid in HCT- 116 DO cells.
1 ver model, we investigated the effect	of	toremifene on the elimination of an
1 ted the tumor growth-promoting effect	of	TP as shown by 81% and 80% tumor-re
1 enhanced the growth inhibitory effect	of	vindesine on both H69/VDS (x 12.0)
1 sed the cancer chemopreventive effects	of	water extract of green tea (WEGT) a
1 tumorigenesis, the protective effects	of	X-rays or heat treatment which caus
2 s arise by a mechanism of double loss	of	a tumoursuppressor gene on 3p, non-c
2 he Walker 256 carcinosarcoma some loss	of	antitumor activity was found with b
2 With the L1210 murine leukemia no loss	of	antitumor activity was found for an
2 s occurred which resulted in the loss	of	ASA as a function of the time period
2 tearylacetamide. Both the initial loss	of	ASA and the increase in stability d
2 om 55 patients were analysed for loss	of	chromosomal heterozygosity using 46
2 ABSTRAC. Loss	of	heterozygosity occurring on various
2 breast cancers were examined for loss	of	heterozygosity (LOH) at tumour suppr
2 rs from female B6C3F1 mice for losses	of	heterozygosity (LOH) at markers nea
2 markers covering 15 chromosomes. Loss	of	heterozygosity (LOH) and/or rearran
2 t aL, we only found tumours with loss	of	heterozygosity in these authors' cle
2 gical types in the prevalence of loss	of	heterozygosity at any locus. There
2 omas and 26 adenocarcinomas) for loss	of	heterozygosity at the pS3, Rb, APC,
2 without such losses, we compared loss	of	heterozygosity data from 51 t~umours
2 ymorphic sites within the genes. Loss	of	heterozygosity occurred in 55% of i
2 creatic cancer, and suggest that loss	of	its regulatory functions may constit
2 some 3. Mitotic recombination or loss	of	one chromosome 3 homologue followed
2 e accompanied by a dose-dependent loss	of	plasma volume. Endothelin-l (I nmol
3 l cycle and a reduction of the number	of	cells in S phase. In contrast, 8-ch
3 he proportion and the absolute number	of	cells positive for the tumor-associa
3 in decisions to intervene. The number	of	days on which moderate hypoglycaemia
3 retic analysis showed that the number	of	disappearing cellular proteins was
3 ption was not influenced by the number	of	donors or patients attending the cli
3 mas has been associated with a number	of	genetic alterations involving chrom
3 a 3-fold increase in the total number	of	hepatic adenomas and carcinomas per
3 gic evaluation of animals at a number	of	intermittent times for the purposes
3 th LLC- IL2 mixed with the same number	of	LLC-IU cells was more suppressive t
3 sufficient to attain maximal numbers	of	lung tumors than that needed for a
3 ficant antitumor activity in a number	of	murine and human tumor-model system
3 a 13-fold increase in the total number	of	pulmonary adenomas and carcinomas p
3 ere was a transitory increased number	of	S- phase heptocytes observed at the

3 ssed in this cell strain and a number of similarly derived normal mammary ep
3 n one and 7 in exon 2. The low number of spontaneous tumors available in thi
3 y abnormal and there are large numbers of structurally abnormal chromosomes.
3 proteins was preater than the number of the newly appeanng ones after the c
3 were evident by a decrease in number of tumors and the percentage of mice w
4 tionalized by postulating the presence of 5HT3 receptors on afferent nerves w

4 have investigated whether the presence of a DNA repair enzyme, 06-methylguani
4 ence in this study shows the presence of a retrovirus in chinook salmon with
4 000 Da, as indicated by the presence of a single band on SDS-PAGE. Amino ac
4 sure), were examined for the presence of activated ras proto- oncogenes. DNA
4 n nonnal plasma (also in the presence of antibodies against tissue factor) o
4 DNA histograms indicate the presence of cell populations with small net qua
4 ctron microscopy confirms the presence of cell- surface microvilli and interce
4 action of 4, X=H, R=Me in the presence of esterase and H218O, did not contain
4 m Pseudomonas sp. BN9 in the presence of glutathione. In the absence of glut
4 oma were also studied hr the presence of HPV by in situ hybridization using
4 ed of plasminogen, or in the presence of I mM tranexamic acid. Plasmin gene
4 ere studied in ELISA for the presence of IgA and IgG antibodies to 5 previou
4 gnificantly decreased in the presence of Li+, Ni7+, Mg2+, Zn2+ or I . Urease
4 h]anthracene with DNA in the presence of liver microsomes from Aroclor 1254
4 els in rats as well as in the presence of maternal toxicity in mice and rats.
4 y, but not dependent on, the presence of Na2EDTA, DL-dithiothreitol (~ 01 t
4 tro we first established the presence of only wild-type p53 and lack of any
4 mine the lung tumors for the presence of other transforming genes. At presen
4 e to generate plasmin in the presence of plasminogen. These cells have been
4 erephthaloyl chloride in the presence of pure dimethyl-°-cyclodextrin (DM°CD)
4 found to be dependent on the presence of pyrazinethiol.