Collocations in science writing

Christopher Gledhill

▶ **To cite this version:**

Christopher Gledhill. Collocations in science writing. Gunter Narr Verlag, 22, 270 pp., 2000, Language in Performance Series, 3-8233-4945-7. hal-01219992

HAL Id: hal-01219992
https://u-paris.hal.science/hal-01219992

Submitted on 28 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Collocations in Science Writing

Christopher Gledhill (2000).

Language in Performance Series No. 22,
Tübingen: Gunter Narr Verlag

270pp.

ISBN 3-8233-4945-7.

Language in Performance Series No. 22, Tübingen, Gunter Narr Verlag, 270pp.

Preface.

This book is based on my doctoral research (1992-1995). It was motivated by a desire to reach out from the *Language Studies Unit* (Aston) and talk with the people in the labs opposite. The book is dedicated to the hard work of the cancer researchers at Aston and Birmingham Universities: Dominique Armspach-Young, William Fraser, Sally Freeman, John Gardiner, Andy Genscher, Helen Mulligan, William Irwin, Peter Lambert, Richard Lewis, Peter R. Lowe, David Poyner, Michael Tisdale, Yaruko Wang and Richard Wheelhouse. They all enthusiastically participated in the survey and were kind enough to allow me to use their publications in my text corpus.

The research presented here was inspired by the work of numerous linguists at Birmingham University, some of whom developed the very first computer-based analyses of texts. At the time I completed the thesis, there were no introductory books on corpus linguistics, large teams of lexicographers were needed to create a 20 million word corpus, and there were no major collections of specialist texts. The situation has evolved considerably since then, although specialist corpora are still rare. At 500 000 words (including 150 research articles), the corpus I use in this book is still a reasonable size, at least for the moment. Phraseology is one of the most exciting branches of linguistics to be involved in at the present time, especially in the fields of discourse and genre analysis. I hope that this book will inspire further work in this particular area.

I would like to extend my thanks to all family, friends and fellow linguists whose help and ideas have helped me with my work, especially Denis Ager, Chris Beedham, Meriel Bloor, Malcolm Coulthard, Beverly Derewianka, Tony Dudley-Evans, Noel and Janet Gledhill, Gill Francis, Liu Haitao, Tim Johns, R. A. (Tony) Lodge, Jacky Martin, Céline Montibeller, Rainer Schulze, Christina Schäffner, Peter Roe, Jean-Pierre Vidalenc and David and Jane Willis. I would also like to thank Mike Hoey, Frank Knowles, Patricia Thomas and John Sinclair as well as the two anonymous readers who kindly read the manuscript and suggested ideas at various stages. They are not responsible for any errors and omissions. Mike Scott at Liverpool University deserves my particular thanks as he introduced me to text analysis by *Microconcord* and *Wordlist* (his program *Wordsmith* has now replaced these programs and is available from Oxford University Press). Above all, I would like to thank Tom Bloor, my teacher and supervisor, for his ideas and suggestions on the final book. His good-natured intellectual rigour has enhanced and encouraged the work of the many linguists who have graduated from Aston over the years.

Christopher Gledhill (2000). *Collocations in Science Writing.*

CONTENTS

Language in Performance Series No. 22, Tübingen, Gunter Narr Verlag, 270pp.

Christopher Gledhill (2000). *Collocations in Science Writing.*

I. Introduction

1. Aims

The aim of this book is to explore the language of science writing. The method is to describe scientific research articles on the basis of a computer-held text archive (a corpus). While many features of language have been identified in scientific texts, I examine one phenomenon in particular: collocation. Collocation is a process by which words combine into larger chunks of expression. Some collocations involve words which seldom occur in other combinations (for example: 'auburn hair', 'rancid butter', 'ups and downs'). Others are turns of phrase made up of words that commonly occur in many combinations ('of course', 'so be it', 'as a matter of fact'). These expressions are all related in phraseology, roughly defined here as 'the preferred way of saying things in a particular discourse' (a formula adapted from Kennedy 1984). My use of the term differs from lexicologists such as Dobrovol'skij (1992) and Howarth (1998). The notion comes instead from recent research in discourse analysis (Moon 1998a and 1998b) and happens to correspond to the everyday use of the term in English to denote skilful mastery of linguistic formulations (e.g. 'in the phraseology of diplomatic circles'). Whatever words we use to talk about these expressions, it is clear they are a key part of the writing process, and it is impossible for a writer to be fluent without a thorough knowledge of the phraseology of the particular field he or she is writing in.

The more specific aim of this book is to demonstrate the role of collocations in scientific English. Although much research has been carried out to establish the range of these expressions in English and in other languages, there remains a great deal to be said about the phraseology of science, in particular the differences between the typical collocations of the language as a whole and the kinds of expressions that are used in very specialist writing. Intuitively, most English speakers are able to guess that expressions such as 'ups and downs' and 'so be it' are rare in science writing. Some expressions or words are seen as more central or stylistically typical in the language than others, a concept critical to vocabulary studies and known as centrality (Carter 1998). What distinguishes scientific English from other

varieties of the language is that it is devoid of such idiomatic expressions. This appears to be a property it shares with informational and administrative prose. These texts are said to be restricted to a limited 'neutral' style. Some linguists identify parts of the grammatical system such as the passive as more typical of science writing, and from this claim that science writing is a restricted form of the general language (or 'sublanguage'). Others concentrate on terminology and point to the processes of naming terms in different specialisms: for them, terminology is central to scientific activity and style is not an issue of importance. Both of these approaches imply that science writing uses a selection of pared-down, neutral features of the language.

In this book I intend to demonstrate that science writing is not style-less and neutral, and that while scientific texts may be devoid of traditional idioms, they employ a system of expression which is as 'idiomatic' (i.e. distinctively fluent) as any other discourse. Most speakers are familiar with the stereotypical features of specialised science writing. For example, verbs are expressed in the passive (*the thermostat beaker was filled with the buffer solution, CoA-transferase brains were homogenized in 10-mM-Tris*) and the text is strewn with arcane symbols and terminology (ranging from the rather poetic technical verb *elute, eluted, eluting, elution* to compound nominals such as *adipose tissue lipoprotein lipase* and *2,2',5'-Trihydroxy-4,5-methylenedioxybiphenyl...*). While these are of course typical and obvious features of specialised scientific language, I explore the extent to which science writing has evolved its own distinct phraseology. The following sample (from a paper published in *Tetrahedron Letters*) demonstrates the problems involved in how we describe science writing:

> Although there are several procedures for the preparation of chiral pyrrolidines and pyrrolidinomes, the majority of these exhibit poor enantiomeric excesses, lack versatility, suffer low yields or some combination thereof. Herein, we describe an efficient asymmetric system of substituted pyrrolidines and pyrrolidinomes that should find general applicability to a variety of modern synthetic challenges. (J. Gardiner, 1992 'Total synthesis of Didehydrodideoxythymidine d4T').

This text has some predictable features of scientific prose and at the same time has a very distinctive style that one would not necessarily associate with science writing, or even with natural, well-formed English. The cohesive devices *thereof* and *herein* strike the reader as archaic or legalistic rather than technical, while some perfectly recognisable English words have taken on a specialized meaning in novel combinations (*exhibit excesses, lack versatility, suffer low yields, find general applicability*). It is clear that even this short

extract is made up of a mix of different styles (technical, archaic, legal, expository) and makes use at the same time of a unique adaptation of the normal collocations of English. The differences in style run much deeper therefore than the usual emphasis on technical terms and verb forms might suggest. The English of science not only undergoes a shift in vocabulary and grammar but also in its discourse features and phraseology.

One particular aim of this book is to demonstrate that there are consistent differences between the collocations of General English and Scientific English, a feature that is sometimes forgotten when science writing is simply seen as a limited grammar or a text dominated by technical terms. Another specific goal is to establish the phraseology of different parts of the scientific text (the Title, the Abstract and so on), and also to establish how far they are stable across a series of different texts with different authors. While technical authors are often assumed to write in a standard formal style that extends across a variety of types of English, the analysis of collocations may reveal much deeper tendencies that are particular to the research article genre. Collocations are symptomatic of strong conventions in specialist writing, although the means by which they become established are difficult to explain. For example, it is highly unlikely that the author of the sample above had to explicitly learn that the expression *suffer few yields* is an acceptable combination in his field. Nevertheless, such phraseological knowledge must be acquired at some stage for the expression to be used across the corpus, in a variety of specialist texts on chemistry. In the survey I carry out in this book, it emerges that scientists are rarely aware of how consistent their phraseology is, although they are concerned with other features of their language.

While collocational patterns are not often consciously identified by individual writers, they are relatively easy to demonstrate on the basis of a computer-held corpus. However, one of the more difficult issues raised in this book is the function of collocational expressions in the scientific text as a whole and in the scientific community at large. Linguists such as Stubbs (1996) have noted that a choice of expression often reveals a rhetorical or ideological stance, and this is an important issue in the analysis of scientific texts. For example in journalism people with cancer can be referred to either as a *patient*, a *sufferer* or a *victim*. In more technical writing, the scientist distinguishes between *patients*, *controls* and *subjects*. And more fundamentally, if there is a consistent phraseology of science writing, one might wonder what purpose it serves in the practice of science, and what relation exists between the language of science and the underlying ideology of science writing. The perspective I wish to explore in this book not only identifies the typical way of saying things but also places these expressions in relation to each other in terms of values. I shall argue that while collocations

are useful units of expression, their relative value depends on their position within the overall phraseological system. The use of the passive voice and technical terms implies certain belief systems that are perpetuated in science writing, and I hope to be able to put these systems in context from a phraseological perspective.

Throughout this book, I wish to pursue three basic research aims. The first is a practical one: to provide a method of describing language in a reliable and objective manner. This is mainly achieved by the use of a computer-held archive of texts (the corpus) collected specifically for the purpose of linguistic analysis, and also by the use of software which calculates word frequencies (the wordlist program) and collects word patterns (a concordancer). However, I also try to demonstrate that the specialist corpus requires a contextual basis, in particular one that takes account of the processes of production of the corpus (as the property of a community of scientists, as well as a text in relation to other scientific texts). Thus while the methodology of this book follows the corpus linguistic approach of Sinclair (1991), its theoretical basis also draws on theories of discourse and genre - especially those of Halliday (1985) and Swales (1990). The practical applications of such a method include the well-documented ability to use the corpus as a tool for language teaching, as well as the possibility of using a corpus as an editing tool and as a source of specialist information. One simple application was suggested by one of my specialist informants: he wanted to know what information to include in Abstracts and how to express himself when writing them, because he felt that he needed to follow accepted practice. Although the field I have chosen is very highly specialised, I also wish to demonstrate that the methodology is sound and applicable to other specialist genres.

The second aim of this book is a theoretical one: to establish a notion of collocation within a theory of language, in particular to discuss the role of collocations within texts. While collocations have become a central issue in the study of vocabulary and lexicology (Carter 1998), their role in discourse and genre analysis has not yet been fully explored. Although many studies conceive of collocations as lexical units which are self contained, with a grammatical structure dependent on one lexical item – i.e. less restricted forms of idioms, a number of studies have emerged recently in which the collocational properties of words are seen as parts of a wider system (for example, Francis 1993, Hunston and Francis 1998). It is possible to list the collocational properties of words in corpus analysis, but it is also necessary to explain how these expressions are related to each other in a particular language or discourse. I intend to demonstrate that while science writing may be very heavily constrained in certain respects, it also allows for considerable

choice of expression. This system of choice appears to be an important aspect of the discourse of science, and a discussion of choice is seen as relevant to the theory of language in general (McCarthy 1984, Halliday 1991).

The third aim of this book is more methodological. I hope to refine certain practices in corpus linguistics, notably by designing a corpus on the basis of a specific discourse community (the Pharmaceutical Sciences Department at Aston University) but also by reviewing the methods by which collocations are identified in texts. The latter is particularly necessary, because at present – and despite the widespread use of the term in many works based on corpus analysis – there is no clear notational convention for symbolising instances of collocation. In order to simplify matters, I use a triangular bracket convention < > for statistical collocations (the node and its collocates identified by word list programs) and a curly bracket convention { } for lexical clusters (families of words or phrases usually present in the context of a word and often with similar meanings). In concordances, node words are signalled in bold, while collocates are underlined. More fundamentally, although most collocational analysis is usually based on the patterns of lexical words (content words), I consider grammatical items to be central to the phraseology of my corpus. Grammatical items enter into collocational relations with longer phrases (a process similar to 'colligation', discussed below) and also form collocational patterns amongst themselves (as shown by Renouf and Sinclair 1991). While the fundamental phraseology of the corpus is revealed by statistical analysis, my analysis depends on a further layer of interpretation. I argue that it is necessary to relate superficial collocational patterns to the general phraseology of the text, most notably by invoking a system of alternative expressions and grammatical metaphor (Halliday 1998). I aim to show that this contributes to a more sophisticated means of conducting corpus analysis, in which the textual properties of collocational patterns are more carefully related.

2. Underlying Assumptions

This book belongs to the British tradition of applied linguistics. Theoretical linguists are preoccupied with symmetry and structure in language. They describe systems of sound, networks of meaning or models of syntax. In contrast, applied linguists attempt to relate theories of language to other fields with the aim of bringing fresh insights back into the discipline. Applied linguistics is not about avoiding theory however; it is about testing theoretical models and engaging with the practical and political problems surrounding language and discourse in areas such as industry, commerce and education.

Applied linguistics involves research in first and second language learning and acquisition, translation, dictionary-building, the study of terminology and specialist languages as well as the critical description of political, administrative and scientific discourse. Work in applied linguistics also tends to address contemporary language. Applied linguists tend to allow linguistic models to emerge through the discussion of data rather than to present the model as the main object of enquiry. This preoccupation with data is often interpreted as 'stamp collecting', but I hope to show here that a useful model of language can emerge dialectically, through the gradual process of demonstration and discussion of examples.

The work presented here has been particularly influenced by the research of applied linguists based at British universities (sometimes known as the Birmingham school, but also as the neo-Firthian school because of the influence of J. R. Firth). This includes the work of J. Sinclair on the computational analysis of language, but also that of G. Francis, S. Hunston, T. Johns, R. Moon and D. Willis on lexical patterns, and T. Bloor, M. Coulthard and T. Dudley–Evans on specific varieties of English. The term 'neo-Firthian' implies a wider group than this (M. Halliday, J. Swales, G. Myers, M. Hoey, M.Stubbs, P. Meara, M. McCarthy, and others). While their work is very often diverse, a number of common concerns have emerged:

- An interest in discourse (language in action, language in relation to its users).
- An emphasis on the close relationship between vocabulary and grammar.
- A preoccupation with authentic non-invented data.
- A preference for computers in the analysis of large archives of language.

In section II these themes are investigated in a review of traditional and applied theories of the language of science. Section III then explores Halliday's notion of lexico-grammar and sets out the design criteria of the text corpus. Section IV then provides a statistical and linguistic analysis of the corpus. This leads me to discuss in section V the implications of a phraseological approach to genre and discourse analysis in general.

Rather than build a general corpus of scientific texts I have opted to focus on the language of cancer research. Over the period of my doctoral research (1992-1995), I conducted a survey of pharmacologists and cancer researchers at Aston University, in Birmingham (UK). There are five main reasons for selecting cancer research as a corpus topic and the group at Aston in particular:

- Cancer research is possibly one of world's biggest medical research activities, served by a large selection of the most prestigious scientific journals.

- Cancer is one of the most emotively reported and well-documented diseases in the popular press. The discourse of cancer research is key to understanding the relationship between the reporting of a scientific breakthrough in the technical literature and its wider reporting in journalism. The fact that cancer is an important topic in public discourse should be justification itself for our attention.
- The field offers an interesting insight into the relationship between language and science. Cancer research articles are written in a very highly refined English. The writing is integrated into a high degree of abstract pharmaceutical knowledge with a complex graphic system of communication.
- Cancer is not a narrow specialism or a single research application but instead involves a broad sweep of activities ranging from theoretical chemistry to organisation management (biology, chemistry, drug synthesis, genetics, patient care).
- The cancer research department at Aston is an important research centre for the U.K. serving the National Cancer Institute (the British version, also based in the region) and it has an above-average output of research with a number of high profile breakthroughs reported in the media over the 1990s. As such it offers an ideal context for a discussion of cancer research writing.

Even within this very specific field, the complexity and degree of specialisation involved in cancer studies means that the corpus would be meaningless without an account of its context. The corpus in turn must represent a reasonably homogeneous linguistic community. The specific linguistic practices of a professional group are at the heart of the genre analysis approach (Swales 1990), although they have received little attention in mainstream corpus linguistics. On the other hand, genre analysis has only recently begun to use computer-based corpora. My hypothesis is that any distinctive 'style' or phraseology I discover can be attributed to a broad community of scientists in pharmacology and cancer research and contribute to a description of the research article genre. Section II in particular explores these themes and discusses in detail the context of the corpus.

3. Definitions of Collocation

A collocation is a familiar recurrent expression. For many linguists, collocations are related to a range of commonly recognised multi-word phrases in language, including catchphrases, clichés, fixed expressions, formulae, free and bound collocations, idioms, lexical phrases, turns-of-phrase and so on. Collocation has been defined in various ways, and definitions depend on the specific aims of the observer. Phraseologists and

dictionary makers, for example, examine the way lexical words behave in certain combinations. The adjectives *strong* and *powerful* can thus be seen to have a similar meaning but a different range of use with certain nouns: *strong argument, powerful argument* versus *strong tea / *powerful tea, *strong car / powerful car*. Once such a restriction is identified for a pair of words, we are dealing with some form of collocation.

However, as the word 'familiar' suggests in my working definition, there is more to collocation than the combination of two or more words. In the following discussion, I attempt to synthesise three different ways of categorising and defining the notion of collocation: Halliday's *statistical / textual* view, the *semantic / syntactic* tradition in lexicology, and the *discoursal / rhetorical* model from discourse analysis. I then go on to propose an overall model of phraseology which serves as a basis for the analysis carried out in the rest of the book. In the corpus analysis sections of this book, Halliday's statistical definition is specifically taken as the first and simplest stage of my analysis, but is then supplemented by further stages of interpretation in order to determine the structural and rhetorical significance of the collocations identified in the corpus.

From a **statistical / textual** perspective, it is generally agreed that no one linguistic definition of collocation is entirely reliable when it comes to finding expressions systematically in large numbers of texts. For this practical reason, collocations have often been defined statistically in corpus-based studies, especially if the analyst is attempting to find examples of typical style. The first stage of analysis to be used in this book therefore follows Halliday, who frames collocation in terms of statistical probability and co-occurrence:

> Collocation is the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur at *n* removes (a distance of *n* lexical items) from an item *x*, the items *a, b, c* .... Any given item thus enters into a range of collocation, the items with which it is collocated being ranged from more to less probable. (Halliday 1961:276).

Van Roey summarises this view in terms of expression or 'usage':

> [collocation is] that linguistic phenomenon whereby a given vocabulary item prefers the company of another item rather than its 'synonyms' because of constraints which are not on the level of syntax or conceptual meaning but on that of usage. (van Roey 1990:46).

A collocate can thus simply be seen as any word which co-occurs within an arbitrarily determined distance or *span* of a central word or *node*. Collocation is thus considered to be the frequency with which collocates co-occur with one node relative to their frequency of collocation with other nodes. From the point of view of many corpus linguists, all that separates collocation from mere word co-occurrence is the statistical level at which the researcher is happy to say that the co-occurrence is not accidental. This approach is also 'textual' in that it relies solely on the ability of the computer program to analyse large amounts of computer-readable texts. Sinclair (1991:68) shows this by noting that the independent probability of *'set'* collocating with *'off'* in the Cobuild corpus is just one in a million (1 855 instances of *'set'* multiplied by 556 instances of *'off'* from a total of 7.3 million words). Yet the actual frequency of collocation is around 550 instances (that is: 70 in a million). The expression *'set off'* can thus be considered a significant collocation without considering other semantic or lexical considerations (1987b:153).

This perspective essentially emphasises collocation as co-occurrence (words which frequently combine) and recurrence (combinations which frequently occur in language). The notion of statistical collocation is integral to Halliday's theory of discourse and the theory is discussed in section III. It is sufficient to note here that a statistical view of language allows the linguist to identify patterns that would not normally be recognised using traditional categories. The textual view of collocation also emphasises the fact that collocations are not disembodied lexical units inserted into the body of a text without modification, but are the result of reformulations and paraphrases which have developed throughout the length of a text. A textual collocation is likely to have a specific textual function or may occur in a rather restricted set of contexts. These expressions can be seen to be couched seamlessly in the surrounding text, and in many of the examples we see below, the collocational patterns of a specific phrase are motivated or triggered by other phrases which appear to be at some distance (a phenomenon observed by Phillips 1985 and Hoey 1991). This is what is meant by 'long-range collocation'.

In contrast, the **semantic / syntactic** tradition defines collocation as a more abstract relationship between words, without reference to frequency of occurrence or probability, shifting the emphasis therefore from the textual co-occurrence of an expression to its potential for lexical combinability. While Halliday's approach to collocation is appropriate to a discussion of discourse and register, style is not the main concern in lexicology. Instead the emphasis is on dictionary making and terminology, and collocations are typically seen

either as units of meaning (lexical items or idioms) or units of grammar (phrases). It is for this reason that collocation is usually seen as a rather restricted category of expression and is also typically limited to the lexical relation between content words. The standard definition is given by Benson:

> Collocations … are fixed recurrent combinations of words in which each word basically retains its meaning. (Benson 1989:85).

Howarth (1996) has presented a synthesis of the mainstream ideas of lexicology and phraseology studies, taking particular account of the Russian perspective (Dobrovol'skij 1992). He notes that the 'composite unit' is traditionally classified according to two measures (1996: 36-46):

> 'Commutability' - The extent to which the elements in the expression can be replaced or moved. As in the free collocation *make a decision* where *make* can be replaced by a series of de-lexical verbs *reach, take* etc., while in the restricted collocation *shrug one's shoulders* there is no alternative to the verb *shrug*.
>
> 'Motivation' - The extent to which the semantic origin of the expression is identifiable, as in the figurative idiom *move the goalposts* [to change the required conditions for success], as opposed to the opaque idiom *shoot the breeze* [to chatter].

Fixed expressions are characterised by the relationship between their component words and the overall meaning of the phrase. Cruse (1986) thus distinguishes collocation as 'syntagmatically simple' i.e. an expression composed of one word in its normal sense with another restricted word (as in: *table a resolution, tender one's resignation*) and idiom as 'semantically simple' i.e. as a single choice of meaning with an unpredictable or non-compositional sequence of words (*let the cat out of the bag, spill the beans*). In Howarth's lexical continuum model (1996:32-33), collocations are placed on a sliding scale of meaning and form from relatively unrestricted (collocations) to highly fixed (idioms):

Free collocation	*blow a trumpet*	'to play the trumpet'
Restricted collocation	*blow a fuse*	'to destroy a fuse', or (idiomatic) 'get angry'
Figurative idiom	*blow your own trumpet*	'to boast, sell oneself excessively'
Pure idiom	*blow the gaff*	'to reveal a concealed truth'

The problem commonly encountered with these classifications (as can be seen in the ambiguous example of *to blow a fuse*) is that is difficult to determine what is meant by 'syntactically fixed', 'unmotivated' or 'opaque'.

In addition to the notion of the collocational continuum, one of the most influential ideas to emerge from the field of lexicography involves Mel'□uk'sč theory of lexical functions. Mel'čuk defines collocation as an semantic function operating between two or more words in which one of the words keeps its 'normal' meaning (Mel'čuk 1995:182). Fontenelle explains this abstract relationship:

> […] the concept of collocation is independent of grammatical categories: the relationship which holds between the verb *argue* and the adverb *strongly* is the same as that holding between the noun *argument* and the adjective *strong*. (Fontenelle 1994:43).

For example, several restricted collocations in English have the abstract function of 'intensifier' (coded by Mel'čuk as '*magn*'): *stark naked, utter foolishness, piping hot*. The vocabulary as a whole is therefore organised into a grammar of intensity, of quantity (*a speck of dust, a pride of lions*), of operation (*to lend support, to deal a blow*), of function (*war is raging, silence reigns*) and so on (Mel'čuk 1998:36-41). By bringing disparate collocational patterns into a broad theory of meaning, Mel'čuk has argued for a universal typology of lexical functions which are realised by a delimited number of underlying lexical functions in English and other languages.

In lexicology and phraseology studies, idioms are seen as the prime examples of semantic and syntagmatic units, and have a correspondingly privileged status (Howarth 1998:169). On the other hand, collocations emerge as less tidy and easy to categorise, being seen as increasingly less fixed and also more diffuse – largely of course because they are often defined in terms that make idioms generally appear to be ideal units. Collocations also tend to be defined as a subcategory of other items. Mel'čuk, for example, sees them as a very specific category: 'Collocations – no matter how one understands them – are a subclass of what are known as *set phrases*' (Mel'čuk 1998:23). Approaching the issue from a different perspective, van der Wouden (1997) has argued that collocation should be seen as the central term in lexicology. He points out that regardless of the way collocations are defined, analysts find more instances of collocation than of idiom in actual texts, and proposes that the notion of 'collocability' requires better definition than the more peripheral idea of 'idiomaticity'. Like many linguists in the generative field (for example, Abeillé 1995), he sees syntagmatic variability as key to the notion of a fixed expression, and suggests that many features of language are idiomatic in this sense:

> I will use the term COLLOCATION as the most general term to refer to all types of fixed combinations of lexical items. In this view idioms are a special subclass of collocations, to wit those collocations with a non-compositional, or opaque semantics. An idiom might even be defined as any grammatical form whose meaning is not deducible from its structure. In this view all morphemes are idioms. (van der Wouden 1997:9).

Makkai (1992) has similarly argued that collocations and idioms can be seen as extended forms of words. Kjellmer makes a similar point:

> Highly distinctive collocations behave in important respects like one-word lexemes. They are often semantically identical or almost identical with single words. (Kjellmer 1984)

Van der Wouden further makes the point that idioms and collocations share a number of properties, not least of which the ability to contain analogies which are not carried on into the rest of the language system:

> [...] you cannot predict that the meaning of *sleep like a log* will denote an intense form of sleeping, but after you have learned what it means, you see that *like a log* is an intensifier. The essence of collocation is that the assignment of *like a log* to the meaning 'very' does not feed other combinations. So even though we have a meaning for it, that meaning is only valid in a certain collocation [...] (van der Wouden 1997:54-55).

From this discussion, it emerges that the distinction between idiom and collocation is difficult to justify on purely semantic or syntagmatic grounds. Instead, collocation constitutes a general system of abstract relations which underpin much phraseology in the language, and range from relatively free to relatively fixed expression. A different perspective, although still within our 'semantic / syntactic' framework, relates collocational patterns to the wider grammatical system, as in the work of Sinclair (1991). For example, Renouf and Sinclair (1991) have noted that the meaning of a lexical item can be predicted by the presence of grammatical items and the sequence in which they are arranged. Thus in expressions such as *an X of*, X is often a quantity, or in *too Y in the Z*, Y and Z are often time expressions (such sequences are termed collocational frameworks). Louw (1993) has noted that clusters of lexical collocations often share a similar semantic profile or 'semantic prosody'. Thus the NP subjects of the phrasal verb *set in* belong invariably to a semantic field with negative associations (*the bad weather, gangrene, the rot, depression ... sets in*). According to this perspective, the grammatical patterns of co-occurrence are an intrinsic meaning of an expression, and any

item which is inserted into the pattern can be re-interpreted in terms of the existing collocational framework (e.g. *a cacophony of musicians* [collective], *the Labour party have set in* [negative connotation]).

In a large-scale study of verb complementation, Hunston and Francis (1998) similarly make a specific link between the grammatical form of an expression (its underlying word class pattern) and its meaning, claiming that the pattern is part of the meaning of the expression. Hunston and Francis identify a number of collocations which share specific grammatical patterns and yet also display a closely related meaning. Here is one example:

> ...sense and pattern tend to be associated with each other, such that a particular sense of a verb may be identified by its pattern. The verb *recover* has two main senses: 'to get better' following an illness or period of unhappiness, and 'to get back' something that was lost. The first of these senses has the pattern 'V *from* n' (e.g. *He is recovering from a knee injury*) [...] and 'V' (e.g. *It took her three days to recover*), whilst the second has the pattern 'V n' (e.g. *Police... recovered stolen goods*). (Hunston and Francis 1998:51).

This can be seen to be an extension of the general principle of delexicalisation, in which lexical items merge into grammatical forms, effectively becoming grammatical collocations (grammatical words collocating with lexical words). The expressions created by grammatical collocation and colligation depend in turn on a notion of extended meaning, as argued by Renouf (1998). The extended meaning of a word or expression is built up over time by its collocational tendencies within different texts. Thus while lexicologists conceive of collocation as a lexical unit and examine the behaviour of component words within this larger lexical item, Firthian and Hallidayan linguists see collocation as a specific grammatical pattern, associated with a particular meaning. The work of Louw, Renouf, Hunston, Francis and others has been much influenced by Sinclair's notion of the 'idiom principle'. Sinclair (1991) argued that meaning is organised through language not by filling lexical items into grammatical context-free slots, but in a system where structure maps onto meaning very closely. He emphasises the importance of syntagmatic sequences as single functional choices, and argues that neither individual words nor deep syntactic structures correspond to natural choice in language:

> 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 conversation. (Sinclair 1987c: 320)

From the 'semantic / syntactic' perspective, we have seen that the notion of collocation has been extended from traditional restricted collocations and idioms (*curry favour, strike a chord*) to less conventional notions such as grammatical collocation (linking grammatical items with lexical items, as in phrasal verbs *refer to, answer for*) and de-lexical verbs (*have a break, take a decision*). Many of these patterns can be seen to obey underlying lexical relationships. The notion has recently been applied to a much wider category of expression following work in corpus analysis, including semantic prosody (clusters of semantically related words: *push through* [*a reform, a project, a law*...]), collocational frameworks (lexical and grammatical collocation: *not only*... *but also*, *find / make it* [*easy, difficult, hard, impossible*] *to* + clause) and colligation (collocation between grammatical categories, e.g. the set of nouns that can introduce NP complement clauses: *the idea, conviction, belief, thought that*). These patterns demonstrate the close correlation between syntax and semantics and are seen as a confirmation of Halliday's (1985) notion of a lexico-grammar: a theory of lexis and grammar as an interrelated continuum rather than as separate levels.

So far we have seen collocations as 'statistical / textual' co-occurrences on the one hand or as 'semantic / syntactic' patterns on the other. However, it is possible not only to examine the internal syntagmatic properties of an expression, but also the pragmatic role of the expression in text and discourse. A third tendency therefore is to examine collocations in terms of performance, in other words from a **discoursal / rhetorical** point of view. From this perspective, idioms such as *to get the sack, to be fired* can be contrasted stylistically with less marked expressions: *to be dismissed, to lose one's job*. The difference between these expressions lies in their emphasis or rhetorical effect, as Moon (1987) and Fernando (1996) have argued. From a discourse analyst's perspective, Moon feels justified in arguing that syntactic and semantic constraints on fixed expressions are not as important as rhetorical function:

> In general, studies of fixed expressions [...] concentrate on their typological and syntagmatic properties. Attention is given to such things as the degree of their lexical and syntactic frozenness, or their transformation potential: and even the primary characteristic of idioms, their non-compositionality as lexical units, may be seen as a matter of the interpretation of a syntagm. However, it is their paradigmatic properties which are of importance in

> relation to interaction. Fixed expressions represent meaningful choices on the part of the speaker / writer. (Moon 1994:117).

Fillmore and Atkins (1994) and Kay and Fillmore (1999) have similarly questioned the need for a distinction between idiom and collocation on the grounds of syntactic and semantic frozenness. Fillmore, Kay and O'Connor emphasised the fact that collocations are culturally salient items which need to be learnt as part of the language. According to their well-known definition, fixed expressions are:

> […] 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. (Fillmore, Kay and O'Connor 1988:501)

In a similar approach, Pawley and Syder have been influential in the area of language learning theory, and were among the first to emphasise that conversational gambits in natural speech were speech acts organised around fixed expressions of the type *it's easy to talk* (a reprimand for some criticism)*, she's busy right now* (denying access by telephone) and *I thought you'd never ask* (expressing relief after permission has been granted) (1983:307). They pointed out that these expressions are effectively social institutions, and have specific cultural functions in the language:

> A lexicalized sentence stem is a unit of clause length or longer whose grammatical form or lexical context is wholly or largely fixed; its elements form a standard label for a culturally recognized concept, a term in the language. Although lexicalized in this sense, most such units are not true idioms but rather are regular form-meaning pairings. (Pawley and Syder 1983:191-192).

This theme was similarly examined by Yorio, whose analysis of a spoken corpus found few traditional idioms, but instead proposed that sentence stems are key to understanding conventionalised fluency in language. Yorio concludes that grammatical accuracy must be matched by a knowledge of such idiomatic expressions:

> Idiomaticity, or native-like quality in written language, appears to be a property characterized primarily by the presence of collocations and / or sentence stems rather than by actual idioms. [...] [A]lthough fluency is possible without grammatical accuracy, idiomaticity is not. Idiomaticity then becomes an excellent indicator of bilingual system proficiency and, as such, it deserves to be further studied and understood. (Yorio 1989:68)

Nattinger and DeCarrico (1992) examined shorter stretches of language than the sentence stem, and related knowledge of phraseology to a system of rhetorical expressions (1992:22). Following Coulmas (1979), they situated collocations within a continuum of increasing rhetorical force: from low to high impact. Nattinger and DeCarrico identified collocations as unmarked choices of expression '[co-occurring lexical items] that have not been assigned particular pragmatic functions by pragmatic competence' (1992:36). This 'unmarked' sense of the term collocation is an interesting departure from the perspectives we have seen above and clearly delimits the syntagmatic definition of collocation from a discoursal one. Nattinger and DeCarrico then contrast unmarked collocations with lexical phrases, defined as 'marked' collocations, in that they have recognised pragmatic functions. Lexical phrases are split into two groups (1992:38-42):

- Lexical units which do not allow paradigmatic or syntagmatic reformulation: polywords: *for the most part, as it were* and institutionalised phrases *how are you? what, me worry?*
- Grammatical frameworks with both fixed and free features: short range phrasal constraints: *a NP [time] ago*, long range sentence builders: *I think (that) [proposition clause X], the ADJ-er [proposition clause X], the ADJ-er [proposition clause Y]*.

The lexical phrase is proposed as an addition to the traditional distinction between idiom and collocation, and emphasises textual function rather than internal form:

> Lexical phrases are parts of language that often have clearly defined roles in guiding the overall discourse. In particular, they are the primary markers which signal the direction of discourse, whether spoken or written. When they serve as discourse devices, their function is to signal, for instance, whether the information to follow is *in contrast to*, *in addition to* or is *an example of* information that it to proceed. (Nattinger and DeCarrico 1992:60)

According to Winter's (1977) theory of clause relations, information in discourse is frequently managed lexically. Nattinger and DeCarrico show that this operates at a phrasal level by the use of global topic markers (*let's look at*), shifters (*OK, now*) and summarisers (*so then*), as well as at a local level by the use of exemplifiers (*how about X?*), relators (*it has to do with Y*), qualifiers (*the catch is that…*), asides (*where was I?*) and so on. Such expressions are typical of the spoken language, but we see below that science

writing has developed a sophisticated system for similar functions (including asides and topic shifters), albeit with different linguistic expressions. While such features may not be statistically significant across the corpus, and therefore do not usually figure in corpus-based analyses of register, Nattinger and DeCarrico claim that such phraseology has a significant role to play in the rhetorical construction of the text. These claims are supported by related studies on the pragmatic function of idioms in texts (Popiel and McRae 1988, Luzon-Marco 1999)

The 'discourse / rhetorical' approach is not concerned with lexis and grammar as such. Instead, the suggestion is that collocations and idioms can be distinguished on the basis of a rhetorical or textual function (as argued by Nattinger and DeCarrico) or pragmatic marking (as argued by Moon). We have seen above that most idioms - such as *sell like hot cakes* (to sell quickly) and *pull a fast one* (to deceive by stealth) - are more marked stylistically than their typical paraphrases, not just for emphasis, but often with very specific information and a limited context of possible use. Moon has suggested that many such idioms and metaphors are deliberately used in speech and writing to bring in shades of evaluation or judgement in comparison with their unmarked equivalents (thus *the trial progressed at a snail's pace* would signal subjective feeling more explicitly than *the trial progressed slowly*). But as Moon points out, these 'prototypical' idioms are rarely found in authentic texts. In practice, the most commonly recurring expressions are likely to be 'lexical phrases' or 'sentence stems' and it is worth noting that apart from Nattinger and DeCarrico's work, these have received much less attention from lexicologists.

A normal text rarely moves in a clear-cut way from unmarked to marked expression, with idioms and collocations visibly demarcated. It is more realistic to picture a text as a sequence of different types of discourse signal, and while most of these expressions are idiomatic in that they have specific rhetorical or pragmatic roles to play, they are not marked as such within the normal reading of the text. Thus while lexical phrases may appear to be idioms from a traditional lexicological point of view, in their normal context they are simply part of the accepted phraseology. When something is 'marked' or pragmatically unusual, we can assume that it stands out from the expected style. Indeed, a knowledge of the expected phraseology is central to being able to step out of it in order to create some supplementary rhetorical effect. For example, Pawley and Syder's sentence stems have very specific and sophisticated rhetorical functions in spoken English: they are natural candidates for the category of idiom. But it does not make sense to suggest that they are permanently marked expressions, especially when we consider that they are commonly used in normal spoken discourse.

To give another example, the British English greetings *How do you do?, How are you? How do?, How's it going?, How goes it? Wotcha!* etc. vary from unmarked to marked in different contexts. The native speaker knows the core items (depending on dialect) and knows implicitly their rhetorical value in the phraseological system. *How do you do?* is felt to be the standard prototypical form, but this does not mean that it is the unmarked, neutral choice used in the majority of circumstances. The corollary of this is that prototypical expressions do not correspond to typical expressions. In addition, a notion of what constitutes 'collocation' or 'idiom' may also depend on an appropriate register or style and part of the meaning of an idiomatic phrase is its specific context of use in which it is deemed to be appropriate (a pragmatic dimension rather than a strictly textual one). Thus from a discourse perspective, idioms (as relatively marked expressions) and collocations (as relatively unmarked expressions) might not be fixed categories, but may be perceived differently in different contexts. Collocations can be said to have a less fixed pragmatic set of uses than idioms; while lexical phrases, with their specific rhetorical roles, occupy a position somewhere in-between. From this basic premise, we can postulate a shifting rhetorical continuum between the usual phraseology of collocation and other more unusual expressions (including original expressions which break with collocational convention or stylistically marked idioms belonging to another discourse).

Collocation emerges throughout this discussion as a powerful but also extremely diverse concept. As van der Wouden (1997) notes, the term collocation itself either refers to the abstract relationship between words or the expression as a whole. Nevertheless, it is clear that although there are differences in application and methodology, all of the approaches we have summarised above converge on an important and recognisable phenomenon, the 'familiar recurrent expression'. Instead of arguing the case for one specific viewpoint, I attempt to see each as compatible and relevant at different points in my analysis. Since the main purpose of this book is to analyse a large corpus of texts, I argue below that the 'statistical / textual' perspective is the most appropriate approach to be adopted in the first stages of corpus analysis. However, the 'semantic / syntactic' perspective brings to our analysis of collocation the important notion of the abstract relationship between words, and the idea that the expression exists as a meaningful unit of choice within the grammar. The 'discoursal / rhetorical' view equally informs us of the role that the expression has within a running text and reminds us to interpret the expression as part of a system of stylistic alternatives. Despite differences of methods, each approach leads us to reconsider the relationship

between words within the collocational expression and to revise the traditional notion of phraseology.

I intend to use the term **phraseology** to refer specifically to the rhetorical or pragmatic use of an expression. The term then stands in contrast to Halliday's 'lexico-grammar' which refers strictly to the cline between lexis on the one hand and grammatical systems on the other (Halliday 1985). The term also contrasts with the notion of 'collocational continuum' in lexicology (Howarth 1998), which refers to collocations as they become less like phrases and more like words. The 'discoursal / rhetorical' approach claims that the pragmatic value of a particular expression constitutes an important aspect of a theory of phraseology. However, few studies of idiom or collocation have taken this perspective, and even fewer have attempted to account for systems of phraseology in scientific texts. My assumption in the analysis below is that although my collocational expressions are originally derived from the corpus on a statistical basis, they can be also usefully described in terms of their textual, rhetorical or pragmatic function. Thus a lexico-grammatical analysis of a specific discourse can be supplemented by an analysis of phraseology.

A further issue at this point concerns the notion of grammatical item (a closed class or functional word) and lexical item (an open class or content word). In the corpus analysis below, I suggest that grammatical items are useful starting points for the analysis of longer stretches of collocation and phraseology. We have seen in the discussion above that grammatical items have usually been left out of collocational studies. Many studies of textual collocation such as Phillips (1985) or Smadja (1993) go further and eliminate 'stop-words', largely because grammatical items are too frequent in the corpus and are reasonably thought to 'collocate with anything'. There is also a similar tendency in lexicology, in which grammatical items are usually considered only as collocations of lexical items (as with prepositional and phrasal verbs). However, as mentioned above, important work by corpus linguists such as Hunston and Francis (1998) on the patterns of grammar, and Renouf and Sinclair (1991) on consistent grammatical features of collocation has shown that grammatical items are fundamental to a theory of phraseology. The 'discoursal / rhetorical' approach has also brought into focus many previously ignored combinations of grammatical items which function as recognisable expressions. For example, many of Nattinger and DeCarrico's lexical phrases contain, ironically, very few lexical items: *just because, to be at it, as is, that's it then, it's all over, he's out of it*. These expressions are considered to be lexicalised, although they function more like utterances than single lexical items. Following on from this perspective, the analysis I set out below focuses on grammatical items as the key elements in longer stretches of phraseology. In section III, I specifically address the role

of collocation in specialised texts and set out more fully Halliday's concept of the lexico-grammar.

The notion that grammatical items are closed class words will serve as my basic rule-of-thumb in order to identify these items. However, I also wish to explore the possibility that high frequency items (such as auxiliary verbs *is* and *has*) play an important role in the formation of collocations and fixed expressions, and assume therefore that such high frequency items are for the purposes of my analysis 'grammatical'. This frequency-based approach to lexis is consistent with Sinclair's view, and allows for a more nuanced analysis of words which are often considered to be at the intersection between grammar and lexis.

II. Language and Science

This chapter sets the scene for the corpus design in section III and data analysis in section IV. The aim here is to justify my specific object of enquiry (science writing in cancer research) and my methodology (an approach within discourse analysis). I set out here the theoretical basis for a corpus analysis of cancer research articles. I explain briefly the relationship between science and language from the point of view of terminology and then from linguistics (especially genre analysis). In order to put the research article genre in context, I then discuss a specific discourse community: the Pharmaceutical Sciences Department, Aston University.

The language of science is a fruitful and well-documented area of research, most notably in philosophy, sociology and linguistics. The role of language in science was the object of enquiry of philosophers concerned with hermeneutics and the reflective function of science (Gadamer, Wittgenstein and Foucauld) as well as theories of knowledge and scientific epistemology (Bachelard, Piaget and Kuhn). In sociology there has been much research on the discourse of science in relation to science policy and the public understanding of science. There is particular interest in the ways in which technical issues are affected by economics, politics and personal agendas (Kevles 1995 sets out a comprehensive history of the discourse of cancer research). For the most part, research on science writing in linguistics has been the realm of applied linguistics, in particular the divergent fields of terminology and discourse analysis. The two approaches can be summarised as follows:

1) Terminology centres on the theoretical relationship between the specialist subject and language. The object of enquiry is that of Languages for Special Purposes (LSP), defined in terms of specialist topic rather than style or other linguistic characteristics (Sager et al. 1980, Sager 1990). The field of terminology has a strong rationalist tradition, derived from its origins in the creation of industrial and scientific standards. Terminologists are often scientists themselves, including proof-readers, editors, abstractors, translators, termographers (builders of term banks and indexes) and information scientists (text-engineers).

2) Discourse analysis discusses the activity of science writing and the role of language use among specialists. Applied research on scientific discourse is known as English for Specific Purposes (ESP: Swales 1981b,1990), with the emphasis being on the problems associated with the use of a specific national language (English) in international science. In applied linguistics, ESP and 'English for Academic Purposes' have become widely recognised fields of research, with dedicated academic journals (*English for Specific Purposes, ESPecialist, Fachsprache, Anglais de Spécialité*). Many specialist areas have come under scrutiny, especially in the medical sciences and areas such as doctor-patient dialogue and the popularisation of science. The field has several theoretical traditions, and applications tend to centre on language teaching.

The historical distinctions between terminology and discourse analysis are beyond the scope of this book, but what is of interest here is the way in which language is seen either in relation to the subject matter (the *special* language: a terminological perspective) or in relation to the scientific activity (the *specific* language: a discourse perspective). In the following sections, I explain these two positions.

1. The Terminology of Science

Scientific and technical terminology is often cited as a powerful factor for change in language. To take a basic example, the number of new chemicals created in English (recognised by the international standards organisations such as IUPAC) far outstrips the number of words commonly recognised in the language as a whole. In organic chemistry alone, there are 750 000 compounds and four million standard terms (including affixes) and a further 30 000 terms in inorganic chemistry (Sager et al. 1980:230). This count does not include the many other terms that are created *ad hoc* within texts, as Thomas (1993) points out.

Terminologists create and define specialist terms, most often with legal status, for example in the statutory use of patents. From the point of view of linguistics, the naming of terms is an attempt to fix semantic universals and situate semantic relations within a paradigm or hierarchy. The notion of paradigm distinguishes a *terminology* (a collection of terms related by an underlying system, most usually within a specific discipline) from a dictionary. The technical notion of *term* and its underlying *concept* is therefore distinguished from the lexical *word* or *name*. The key area of terminology however is the definition, 'the verbal description of a concept'

(Picht and Draskau 1985:65). Systems of definitions present a complex area of research and Picht and Draskau summarise the dynamics of definition in terms of internal or external dimensions. Logical definitions of internal or *intensional* characteristics (an entity's shape, colour and other 'independent' properties) can be placed alongside an analogical definition of external characteristics or *extension* (the term's associated purpose or functions) (1985:47). The matter is complicated by the fact that an established concept in one discipline can be interpreted differently in another. For example, the iron chloride molecule $FeCl_3$ is important for electricians as well as textile technologists, but has a different definition (extension) in both fields (Sager et al. 1980:72). As we note in our survey below, biochemists, micro-biologists and pharmacologists have a very different perspectives of the central concept of *cancer*.

Beyond the mechanical stockpiling of terms, the process of creating terminology itself has an impact on the rest of the language system. In a major work on the notion of nomenclature, Cahn (1979) noted that all words in the general language could potentially be pressed into service in science and technology using conventional resources such as conversion. For example the noun *clone* can become a technical verb *to clone* and then be re-introduced into the general language. Scientific derivation also adapts the morphology of the language in order to create subject-specific neologisms. The derivational systems of Greek and Latin are fully employed in English and provide a complex system of fine distinctions. In chemistry the form *-ic* indicates more oxygen bonds, as in *sulphuric acid* (H_2SO_4), and contrasts with *-ate,* used to refer to *sulphate* SO_4 with a valency-2 ion. These can in turn be contrasted with *–ous,* which indicates a decreased number of oxygen bonds as in *sulphurous acid* (H_2SO_3) (Scott 1991:272-278).

Lexical derivation takes the form of compounding, in which words are juxtaposed by leaving a space or hyphen between individual elements. Compounding involves the formation of complex nominals, and this process of term creation has had profound effects on the syntax of English, as noted by Huddleston (1971), Lackstrom et al. (1972, 1973) and more recently by Halliday (1998). Huddleston noted that scientific English has four major nominal categories: adjectival compounds (*compressive force*), verbal nouns (*air-conditioning, town planning*), de-verbal compounds (*dust collection*), and operation compounds (a grammatical reformulation, for example *temperature change* from *a change of temperature*). Sager et al. (1980:268-269) similarly identified the complex semantic interactions between the noun phrase head and its modifier. They established ten dominant categories of lexical collocation in English:

1) head compared with the modifier	*ethane-type interaction.*
2) head made of a specified material	*oil film.*
3) head has a new property	*low octane.*
4) head has a specific use	*cutting tool.*
5) head is associated with its product or origin	*malt beer.*
6) head operates on the modifier:	*enzyme reactivator.*
7) head operates as specified by the modifier	*sliding key.*
8) head is part of the modifier	*pedestal cap.*
9) head is identified by the modifier	*gold standard.*
10) head 'takes place at' the modifier	*cytokine tumour.*

We can see that collocational systems in scientific terminology are particularly complex. Terminologists have demonstrated that there is an underlying grammar at stake in science writing, a view which serves to counteract the folk-view of terminology as simply the classification of terms and taxonomies. However, although this is an important and difficult field of research, terminology still tends to prioritise the complex nature of nominals and lexical collocations. More recent work has however concentrated on semi-technical terms, words such as *analysis, effect, transformation* (Baker, Francis and Tognini-Bonelli 1993), on general words borrowed by hard science such as *charm, strange, up, down* (Pavel 1993 a / b) and the collocational properties of verbs and verb complementation in science writing (Thomas 1993, Pearson 1998). These developments in terminology do not however address the concept of discourse or varying style within the research article genre, since terminology is only concerned with the specialist subject matter. Terminology is essentially about managing the terms and concepts of a scientific discipline, and the issue of style is, perhaps reasonably, a matter of less importance. As a consequence, research in terminology therefore centers on attempts to delimit the 'Language for Special Purposes', either by seeing LSP as a system of terms, or by seeing LSP as a very abstract and specialised language variety.

By limiting the meaning of LSP to a system of terms, Picht and Draskau represent a traditional but also fairly widespread view of language and science. Picht and Draskau see the difference between the LSP and the general language as a continuum of *abstraction*:

> Depending on the pragmatic function and the context of situation, including an epistemological factor, the same topic within a special field lends itself to discussion at different levels of abstraction. (Picht and Draskau 1985:5)

Contrary to the common-sense view that terminology tends to be about 'specificity', Picht and Draskau note that abstraction implies an increased level of conceptual generality. Thus while 'Cologne Cathedral' indicates a specific real world object (denoted by a *name*), the concept CATHEDRAL is abstracted away from outside reference to a generic idea (denoted by a *term*). Abstraction is reflected in the characteristic nominal style of the LSP, while the general language has 'a zero level of abstraction' (following Ure 1971, they claim that this corresponds to a lower lexical density). Picht and Draskau further characterise the LSP as 'monofunctional', in that it cannot be understood by the lay person, is restricted to exclusive groups and is seen as a non-essential variety in the wider community (1985:10-11). The implication of this is that the terminological system is synonymous with the LSP and that the difference between an LSP and an even more abstract *artificial language* (a non-linguistic form of representation involving algebra and chemical formulae) is one of degree. This use of the term LSP is similar to that of *sublanguage*, a concept also originating from the field of terminology (Lehrberger 1982) but also widely used in corpus linguistics (Barnbrook 1996).

However an alternative view has emerged, in which the central concept of the term has been challenged, and the 'special' nature of the LSP has been eroded, largely because of the increasing tendency for sciences to become interdisciplinary. The emphasis has turned instead to 'knowledge-banks' rather than 'term-banks' (Papegaaij and Schubert 1988, Thomas 1993). Many terminologists see the LSP as a variety of the general language, its difference lying in functionality rather than abstraction or degree of specialism.

Following the functional linguists Hjelmslev, Bühler and Halliday, Sager, Dungworth and McDonald (1980) consider the function of terminology and the LSP within a system of discourses. Science writing is defined not just in terms of conceptual abstraction, but in terms of its relation to different types of discourse, and to different structures of knowledge. Firstly, *conceptual discourse* is concerned with reference beyond the environment of the text into the abstract conceptual world of scientific knowledge. *Perceptual discourse* on the other hand, involves reference to the immediate physical and temporal context of the text itself. Finally, *metalinguistic discourse* (including extratextual comment) is said to untypical of scientific text and is a resource that appears to fade away as the language becomes increasingly graphic and conceptual. Sager et al. also make an interesting distinction between the LSP and *register* (in the Hallidayan sense). Halliday uses register to refer to the traditional 'modes of discourse' such as the *language*

of narrative, the language of transaction, the language of exposition which are not types of texts but rhetorical events which emerge in a long stretch of running text or dialogue (1985:318). Sager et al. instead point out that while register is a useful term for forms of interaction between different discourse communities (between journalists and non-journalists, for example), the LSP exists and evolves within the discourse of a specific scientific community (1980:4).

Although terminology is often seen as the analysis of fixed concepts, Sager et al. emphasise the changing and dynamic nature of scientific patterns of thought. Science innovates and forms new paradigms, making a high demand on the terminological resources of language (1980:xviii). They distinguish between conceptualisation, the attempt to fix and define concepts, and reconceptualisation which involves the changing functional perspective of concepts and terms from discipline to discipline and text to text. The term 'sun', for example, is conceptualised differently in different discourses:

a- You can't see that bird because of the sun (*perceptual*)
b- The sun is a star. (*conceptual*)
c- The Germanic word 'sun' is a noun (*metalinguistic*).

Reconceptualisation can also be seen in the changes of expression that take place within the same text. Broadly speaking, this functionalist approach leads to a view of language as not only the encoding of knowledge but as a primary tool in the negotiation of claims and the development of scientific paradigms. From a similar perspective, Béjoint (1988:365) sets out to question the fixedness of terminology and conceptualisation. He inverts the terminologists' traditional metaphor of the 'constellation of concepts' to make the observation that as one's viewpoint changes, so the conceptual constellations undergo a shift in perspective. Béjoint examines the characteristics of scientific and technical words that are often claimed to hold true by terminologists (1988:358):

- Scientific terms follow a chain of definition from LGP words to LSP terms.
- Scientific terms enjoy an absence of ambiguity in context and out of context.
- Scientific terms avoid figurative or metaphorical meanings.
- Scientific terms have origins that can be definitely traced.

Béjoint asks whether such terms as *key idea pointer, bone tissue* or *bacterial culture* can be considered unambiguous out of context, can ever be traced back to original definitions or usages, or can be held as un-metaphorical. Béjoint challenges the underlying assumption that greater precision can be

defined out of context, a point that appears to contradict many scientists, professional translators and terminological commissions (such as the *International Standards Organisation*). His key point, however, is that the process of terminological definition is circular, and this touches at the heart of the rational nature of naming and nomenclature in science. These comments are echoed by Godman and Payne (1981:24), who point out that the very idea of an idealised knowledge structure is exposed to the same flux and uncertainty that is prevalent in the general language. Thus the meaning of a term is dependent on its position relative to other terms and its use in the text, rather than a fixed abstract definition. Béjoint's position is well-known and has led to a greater emphasis on textual evidence in terminology. Thomas (1993) and Pearson (1998) in particular have demonstrated that a corpus of texts is useful in order to gain contextual information about specific terms, a methodology also exploited in experiments with automatic translation (Schubert 1986). Although their aims are different to those pursued in this book (they are interested in the definition or translation of terms rather than the style of science writing), their methods demonstrate that the concept of collocation is more established in terminological work than in other areas of linguistics.

This discussion leads us to examine the scientific text itself and its role in the formation of terminology. The Canadian linguist Pavel (1993 a / b) has emphasised the role of the research article in the formation of terminology. She postulates that terminological change is contrary to stereotypes unplanned and opportunistic, and largely emerges from the processes of scientific writing itself. Other linguists (such as Linstromberg 1991) have noted that metaphor is a key feature of science writing. In addition, Vidalenc (1997) points out that the 'natural language' philosophers preferred simple metaphors such as Aristotle's substitutions and comparisons or Austin's speech acts. Salager-Meyer (1990a:354) argues that metaphors can become dominant in specific research areas. She reports that 70% of head nouns in medical terminology tend to be metaphorical collocations involving structures (*nerve roots, abdominal walls*) while the rest involve processes, functions and relations (*migratory pain, vehicles of infection*). In addition, terminologists such as Koch (1991) and Pavel see the particular choice of a metaphor as vital in the long-term chances of survival of a specific term, a neo-Darwinian notion evoked by such writers as Cavalli-Sforza and Felman (1989) on the cultural evolution of discourse and Chesterman (1997) in his discussion of collocations and memes as translation units.

Pavel specifically examines the effects of interdisciplinary research in the terminology of fractal science. Since fractal imagery is largely adapted as

metaphor from everyday language, its terminology is particularly transparent to non-experts. Pavel and Boileau's (1994) book of fractal terms not only contains definitions but also typical collocations and synonyms of the main entries. Pavel and Boileau thus very clearly identify semantic criteria as consistent features of syntactic patterns (similar to the 'semantic / syntactic' perspective discussed above). For example, compound noun phrases display inclusion (N + N = *particle-cluster*), adjective + noun phrases exhibit gradual 'superordinates' (*chiral chemical compound*), intransitive N + V collocations show specialisation in the verb (*the product crystallises*) and V + N patterns typically display an empirical measure or directionality (*conserve scale*) (1993b: 5). They interpret these patterns as significant constraints in the formation of new terminology, and argue for their inclusion in dictionaries and term-banks. As Béjoint and Thoiron point out, it is more interesting for the non-expert to know the typical processes and agents involved with a certain term than to know which grammatical category it belongs to:

> S'agissant par exemple, du domaine de l'immunologie, il est plus intéressant pour le traducteur ou le rédacteur de connaître les différents acteurs du processus de défense immunitaire, ainsi que leur mode de fonctionnement, que de savoir à quelle catégorie grammaticale ils appartiennent. (1992:8)

Thus the role of the terminologist has moved from providing definitions and basic grammatical features to setting out a phraseology of meaning. Besides constituting patterns of particular importance in the conceptualisation of fractal imagery, Pavel considers the role of these collocations within the text. Her claim is that new formulations effectively reconstruct the terminological knowledge structure of science. As new phrases become neologisms and accepted terms, these in turn bring along their own suite of associated metaphors, sometimes from different disciplines. Pavel refers to these metaphors as *LSP collocations* (1993a:29). She recalls the example of the theatre in one model of artificial intelligence (namely: Schank and Abelson 1977), where terms such as 'scripts', 'actors', 'thematic roles', 'frames' and 'props' help to conceptualise the brain as 'a theater of mental representations' (1993a:25). Such terms not only permit analogy in creating a new conceptual space, but more importantly they bring along the phraseological patterns from their original context. These terms are initiated, negotiated and finally accepted by the wider scientific community:

> ...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. (1993a:29)

Thus fixed collocations are instances of established terminology, to be contrasted with expressions which represent new claims and are more negotiable, or 'up for grabs'. Reversing the process, as scientific metaphors and new collocations (such as 'black hole', 'primal soup', 'gene pool') are disseminated into popular culture, the new term implies an accompanying belief system. This to- and-fro of concepts, with attendant belief structures, is encapsulated by what Pavel terms the *thematic proposition* (1993a:30). The term therefore comes with its own intellectual baggage, and can be seen to infect the knowledge structure of science as well as reflect it:

> ...languages are seen not only as social tools that human communities have created and are continually refining for communication purposes, but also as agents that constantly condition individual behaviour by virtue of social interaction in historically, geographically, and culturally defined settings. (Pavel 1993a:23)

Pavel's empirical and theoretical observations on the lexicon of fractal science are a useful glimpse into the work that has been carried out in the field of terminology. Terms are no longer seen as just highly technical words with fixed meanings. Even in the traditional view, terminology is seen as contingent and dependent on the conventions of specific disciplines. It appears that terms need to be grounded in their subject-specific and textual context just as much as they require precise definition. In addition, general words and fixed phrases can be equally used as specialist terms, and terms can be interchanged between experts and the community at large. Pavel's LSP collocations provide us with a metaphor for expressions with some value: they are created in texts and compete for the attention of readers and scientists. The concept of the collocation also turns out to be a useful intermediary between the word and the text. They also appear to bring along their own conceptual paradigms. The concept of a dynamic terminology therefore provides us with a useful link between the rational approach of terminology and the empirical perspective of discourse analysis.

2. The Discourse of Science

> Even Descartes, that great and passionate advocate of method and certainty, is in all his writings an author who uses the means of rhetoric in a magnificent fashion. There can be no doubt about the fundamental function of rhetoric within social life. But one may go further, in view of the ubiquity of rhetoric, to defend the primordial claims of rhetoric over against

> modern science, remembering that all science that would wish to be of practical usefulness at all is dependent on it. (Gadamer 1976:68)

The terminological approach to language suggests that the way in which a specialist subject matter is reflected in language is central to the understanding of science. The discourse approach leads us in a fundamentally different direction: to examine the relationship between scientific texts and the goals and practices of scientists in their working environment, in other words the discourse of science. The term discourse is used to imply that while style, lexis and grammar are important tangible features of science writing, they also function as pragmatic choices within a specific discourse. The term 'discourse of science' therefore emphasises the role of rhetoric in science and sees linguistic interaction, especially the privileged genre of the research article, as a central mechanism in the development of scientific ideas.

Discourse analysis is concerned with a number of issues, not least of which the means by which texts are formed, and the role texts play within specialist disciplines and in the wider social context. Rather than seeing language as a vehicle for scientific abstractions, discourse analysis views language as a barometer of the social and professional context from which it emerges, changing as the social variables, textual conventions or topic change. Swales (1998) has recently argued that to examine the context of science is to understand the working practices of research, including the world outside the laboratory. Scientific texts are written specifically by scientists interpreting data, attending conferences, submitting articles to refereed journals, keeping up with the specialist literature. But these texts are also ultimately a result of scientific programs of research backed by charities, corporations and governments. Even the most mathematical scientific paper leaves traces of human involvement at every stage of its production and represents thousands of choices of presentation, expression and content. The astronomy journal *Celestial Mechanics*, for example, is dominated by mathematical argumentation and algebraic formulae, punctuated by the occasional '*but'* and '*and also'*. Yet the titles and abstracts in this journal are written in natural English: clearly language has an important persuasive function in the efficient presentation of arguments and data, even where the scientists might claim that 'the facts speak for themselves'.

The context of the scientific text is clearly important, but an emphasis on context still implies that language is peripheral and used in a mechanistic or representational way. The information view of language, posited by rationalist theorists such as Escarpit (1976) implies that language is unchanged from one context to the next: science transcends language, and language simply provides a universal conduit which may be by-passed in

favour of other systems. Language is thus seen as an encoding and decoding device for atomistic information. But this view is incompatible with what we know about written texts in scientific communities. In one of the best known studies of science writing, Latour and Woolgar (1986) demonstrated the subjectivity of science: how scientists need to be persuaded of scientific innovation and were concerned as much with the status and reliability of their informants as with the conceptual validity of their findings. This was the one of first studies to assert the key role of the academic research article in the dissemination of scientific ideas. However, the distorting effect that science has on language is not just evidence of the importance of form over content. Halliday (1998) has argued that scientific activity creates new forms of language over time, and this is necessary in order to express new meanings and to propagate ideas outside the scientific community. Halliday and Martin (1993) have proposed that not only do the social external factors involved in the production of texts have to be taken into consideration, but something of the symbolic (semiotic) status of the text plays a role in the creation of scientific knowledge. This is the approach typically adopted by neo-Firthian linguists in their analysis of scientific texts (including Myers 1990, Ventola 1991, Mauranen 1991, Halliday and Martin 1993). The Firthian approach to language differs from mainstream descriptive linguistics in that it interprets language as a function of society and sees language as fundamental in the construction of human knowledge. This is clearly a model that addresses the concerns of the ESP researcher as well as the terminologist.

In his study of the processes of re-editing in science, Myers (1990) pointed out that in most fields ranging from the philosophy of science, to cultural studies and the sociology of science, there is a constructivist consensus that language or society effectively creates knowledge. From the perspective of epistemology, scientific truth cannot be anything but 'rooted' in its culture, and language is seen to play an important role in framing scientific thought. Relativist and hermeneutic philosophy (Wittgenstein 1957, Heidegger 1966, Gadamer 1976) rejects the idea that language can represent conceptual truth values, instead claiming that knowledge is contingent and subjective within the historical frames of reference of natural language. The natural language philosophers (Austin 1962, Searle 1969 and Grice 1975) also came to reject truth values, and instead established a framework for the fields of pragmatics and discourse analysis (Verschueren 1999). They saw meaning as conventionalised in language rather than referentially encoded in it, and argued that the criterion for good science is not its ability to express truth values but the extent to which it can be understood within natural language. A similar view of language use was elaborated by Lévi-Strauss (1962) and Barthes (1966) in the semiotic construction of social mythology. Semiotics

emerged from Saussure's theory of meaning as a relationship within a structural code rather than as the property of external truth or reality. From this background Foucauld (1972) was to question the way certain areas of science (psychiatry and clinical medicine) regulate knowledge in relation to other disciplines and establish their own coherence as institutions. Importantly, Foucauld saw discourse as central to scientific practice.

If the Firthian linguistic approach shares this perspective, it is in the idea that language is the place not only for the construction of conventional meanings, but also as the medium for the binding of social relations. As Firth says:

> We must apprehend language events in their contexts as shaped by the creative acts of speaking persons. (Firth 1957:190)

While collocation and contextual meaning have been the trademarks of Firth's approach, his ideas have also been influential in theories of scientific text, especially in the work of M. Halliday. Whereas other approaches (cognitive, sociological, ethno-cultural) see language as a reflection of mental processes or social context, Halliday sees discourse as a social context in and of itself. Halliday claims that the influence of scientific writing extends well beyond the confines of discourse communities. He sees science as a discourse which competes with others for attention and dominance in industrialised societies. Halliday and Martin (1993) propose a marxian view of science, characterising scientific discourse as part of an authoritative system of social control, as did Foucauld in his governmentalist theory, as well as many philosophers in the context of science such as Godley, Guba and Lincoln and Saville-Troike. Halliday and Martin have drawn attention to the pervasive effects of scientific practices on our everyday language and to the alienating effect of science on those who have not been trained to handle the discourse. Halliday distances himself however from constuctivism: 'the unreal choice between *language expresses reality* and *language creates reality* (Halliday 1991:59). Instead, language is seen as a scientific tool for getting at reality. His aim is therefore not to deny scientific values, but to decode scientific discourse and make the discourse accessible in education, a goal shared by other neo-Firthian linguists (for example, Drury 1991, Derewianka 1994).

A text is bound therefore to be a discourse, it cannot be disassociated from its context (as in formal grammars) and cannot be considered to be simply a grammatical realisation of a set of propositions (as suggested by textgrammarians such as de Beaugrande and Dressler 1981:89). Halliday emphasises discourse as the product of simultaneous interaction and communication:

> As performers and receivers, we simultaneously both communicate through language and interact through language; and as a necessary condition for both of these we create and recognise discourse... (Halliday 1977:165).

A functionalist account of the language of science does not make a distinction between a 'special language' (LSP) or the general language. The concept of 'special' is seen as questionable, and Halliday refers to the broad category of register as well as 'restricted languages' which appear to have limited social functions (games, greetings, recipes). As far as Halliday and Martin (1993) are concerned, the essential difference is simply between scientific discourse and other, competing discourses, although science writing has a superior cultural position. Similarly, the so-called monofunctional theory (Picht and Draskau 1985), which characterises the LSP as a language of abstraction, falls foul of much research in the context of science. For example, Godley (1993) observes that the terminological system of chemistry is often redundant and arbitrary (not to say ambiguous), with characteristics that differ from one specialism to another and between different countries. In chemistry, for example there is debate about whether metals should be the 'heads' of noun phrases or the other way round (thus meaning that valency is reflected in modifiers). In addition, editors and writers make considerable efforts to explain local conventions and much of the chemical research article (especially Introduction sections) can be seen as a reformulation for the benefit of outsiders. This kind of evidence challenges the image of precision and uniqueness that is imagined in a theory of abstraction. It appears instead to support the observations of Kuhn, Foucauld, Kevles, Knorr-Cetina and others that scientific knowledge is pragmatically conflictual and planned rather than inherently consensual and self-evident.

Discourse analysts therefore reject the term 'special' in LSP, and refer instead to terms such as **variety** (Richards and Schmidt 1983). A variety is commonly seen as a type of language which varies within a general system, and there is no implication that it is limited in function to a specialism or set apart from what is considered to be the general language system. As such it serves as a generic term. Much work on scientific writing however has been conducted on the basis of the LSP (as we have seen in terminology). Other terms have come to be used for specific texts including 'register' (Halliday 1966, Biber 1996), 'genre' (Swales 1990), 'text type' (de Beaugrande and Dressler 1981:85), 'sublanguage' (Lehrberger 1982, McEnery and Wilson 1996) and 'special text unit' (Sager et al. 1980). As might be expected, none of these terms is exactly interchangeable and each carries with it a different view of the relation between the general language and the specific variety.

Sager et al.'s 'special text unit' demonstrates the problems that emerge when linguists attempt to pin down the variable features of texts. In this functionalist model, the primary functions of texts are broken down into categories: status and topic. 'Status' is determined by the knowledge structure which a text aims to represent and modify. 'Aspect' is subcategory of status: the use to which the text is to be put (administrative, pedagogical, descriptive...) (Sager et al. 1980:102). 'Mode' is a also subcategory of 'status', representing formality and planning involved in the text. 'Topic' involves participants' knowledge and level of reference (from specialised to popular) and also includes 'field' (from the very broad field of physics to the narrower field of nuclear physics). Sager et al. (1980:120) claim that these dimensions manifest themselves in various prototypical categories or *special text units*:

- Essay - focuses on the producer's appreciation of reality.
- Schedule - essentially topic-centred and list-like.
- Report - tailored to the receiver's needs.
- Memo - tailored to the receiver's status.
- Dialogue - interactive and flexible.

For Sager et al. (1980:125), texts are primarily categorised according to intentions: informative, evaluative, directive and phatic. Most observers would recognise that purpose accounts for many differences in form. But as with many textual categories devised by linguists, 'special text units' do not correspond to real texts. In reality, there is no way of exclusively fixing a text into one or another category. For example, research articles in particular can be seen to correspond to the first three STUs we see here (essay, schedule and report).

While Sager et al.'s approach provides us with an intuitively symmetrical system, more context-dependent models have been advanced. Swales' theory of genre analysis has been of the more influential models of scientific discourse, based on the early work of Latour and Woolgar and on Bachelard and Foucauld's conceptions of practice in science. Working in English for Specific Purposes (ESP), an area which is largely concerned with training specialists in language teaching (principally in English), Swales (1990) is recognised as a major initiator of ethnographic approaches to the study of specialist discourse.

Swales proposed that the linguist should attend to the practices of the language user, in particular by analysing texts from the point of view of the specialist and by respecting the terms and values of the specialist community. Any text that has a value among the scientists or professional group in

question is termed a genre. The linguistic characteristics of the genre are seen as secondary to its status in relation to other genres and its value depends on the institutional framework of the scientists or specialists concerned. These groups are in turn defined as discourse communities: '...socio-rhetorical networks that form in order to work towards sets of common goals.' (1990:9). Thus while speech communities are defined by the language they speak (with different registers and dialects), discourse communities are defined by what they are talking about (with different genres and jargons). The discourse community always consists of individuals with different interests and specialisms, but the group is also defined by a common aims and the fact that all members are aware of the central issues and debates that preoccupy the community as a whole, even if they do not actually ascribe to them all. Political parties, trade unions, professional associations, commercial companies, government organisations, campaigning lobbies, and voluntary interest groups are therefore all considered to be discourse communities. Successful discourse communities evolve efficient mechanisms of interaction and control. These mechanisms include 'control of technical vocabulary' and the establishment of a professional 'hierarchy of expertise' (Swales 1990:32). The texts used by the group, its genres, are central mechanisms of interaction within the system and are seen as '...the properties of discourse communities... classes of communicative events which typically possess the features of stability, [rhetorical] move recognition and so on.' (1990:9). In other words, a **genre** is a particular language practice, a text type with a variable but implicitly recognised set of linguistic features. Scientific communities recognise a complex system of genres: text books, review articles, peer-review articles, research journals, grant proposals, lab reports, calls for papers, conferences, seminars, newsletters and so on. Unlike other definitions of genre which we encounter below (Biber 1994, for example), Swales' notion of genre implies that there is a discourse community behind it regardless of linguistic or functional definitions of the text.

The language of the genre is seen as very heavily constrained, at least from the point of view of rhetorical structure and effect (Swales places less emphasis on grammar and vocabulary). Swales claims that his analysis of textual genres ultimately stems from Propp's (1928) 'Morphology of the Folktale'. Folktales work because their readers are familiar with conventional rhetorical events, so readers expect *a damsel in distress* (a conventional plot device) or *the couple lived happily ever after* (a conventional ending). The point is that these events have conventionalised (arbitrary) wording, and are highly restricted in content and outcome. Research articles in science have similar devices, which Swales terms 'moves' (described below). Swales thus sees the genre as means to an end, fulfilling a definite set of communicative

purposes (entertaining the audience or selling scientific ideas) and, importantly, owing its existence to a more or less loose set of rhetorical structures and labels which have been agreed by the group (fairy tales with a series of protagonists, review articles with acknowledgements and methods). It is not necessary for the speech community or the discourse community to be consciously aware of the exact linguistic features of the genre, but generally genres are intuitively recognised and agreed concepts. Swales contrasts genre with register (1990:41) which he defines as a linguistic definition of a certain text. According to this view, register is a linguistic category while genre is a social institution.

Swales' approach has been influential, but it is so different from that of other linguists that the basic terminology and the theories underlying the different terms have become confused. The originality of Swales' analysis is that genres are defined in relation to other genres, not just by a series of internal linguistic features or external social functions. This differentiates genre from **sublanguage** used as a textual category by several corpus linguists, including Barnbrook (1996), McEnery and Wilson (1996) and Pearson (1998). As the term sublanguage itself is derived from terminology rather than discourse analysis, many of these works are oriented to a linguistic description of terminology, or tend to analyse very broad categories of text rather than specific text types. Barnbrook (1996: 122) describes a sublanguage as having:

1. limited subject matter.
2. lexical, syntactic and semantic restrictions.
3. 'deviant' rules of grammar.
4. high frequency of certain constructions.
5. unusual features of text structure.
6. the use of special symbols.

This definition combines features of the LSP or 'special language' and the 'artificial language' ('the use of special symbols') as well as bringing other important characteristics into the picture (such as unusual features of text structure and 'deviant' grammar).

Rather confusingly, Swales' view of genre also differs from the work of Biber and Finegan (1994) where the term register is seen as a social convention, and conversely genre is seen as a regular set of inter-related linguistic features. We have also seen that register can be usefully defined as the text types used to communicate between the discourse community and the general speech community, a concept that is more in line with Halliday's view of register discussed below (Sager et al. 1998). Since Biber's concept of

register seems at odds with Halliday's discoursal notion of the term, it is appropriate at this stage to simply adopt Swales' notion of genre and Halliday's concept of register, noting that these terms are used differently outside the field of discourse analysis.

The claim advanced in this book is that discourse analysis provides a more accurate account of the context and grammatical features of language varieties than the register approach adopted elsewhere (Biber 1994, for example). In Swales' analysis, and unlike Biber's (1986) concept of register or Barnbrook's (1996) use of the term 'sublanguage', the principle is that the same grammatical feature may function differently in different contexts. Any evidence to suggest that certain features function differently in the general language and the specialist variety tends to undermine Biber's view of register, which places a high premium on identifying differing distributions of linguistic features and grammatical categories. Biber's 'multifactorial' approach has been to analyse large groups of grammatical features (from a tagged corpus, such as passives and relative clauses) and to correlate their relative frequency with certain intuitive internal functions of the texts involved (such as abstraction, narrative structure). This has led to important work on specialist texts (Biber, Conrad and Reppen 1998). However, this approach does not account for the fact that the same grammatical features may be present in two text corpora but function differently, in which case linguistic cluster analysis is incapable of accounting for these features of the genre. Swales therefore calls to attention the very specific means by which specialist discourse appropriates existing linguistic features and changes their nature. He calls this the discourse coherence of a linguistic feature, and the principle is derived from Firth's theory of meaning.

Swales (1981c) fist demonstrated discourse coherence in his analysis of the past participle in technical English. He found that participles function mostly to bring the reader's attention to non-linguistic text (a table, figure or illustration as in *the curve shown, the list given*) or are used idiomatically as premodifiers (as in *a given reaction*) in a similar way to classifiers as in *a certain reaction*. He argued that these uses are particular to scientific discourse, and have developed a unique function within the research article genre. I have similarly noted (Gledhill 1995b) that numbers are used throughout pharmaceutical research articles as 'pronomials', replacing references to long chemical names. This has consequences for the rest of the pronomial system of the text (especially the range of anaphora, as noted by Liddy et al. 1987), and presumably implies that pronouns have a different profile of use in chemistry texts. These examples certainly fit Barnbrook's description of 'unusual features of text structure' and perhaps also 'lexical,

syntactic and semantic restrictions'. The point is however that in the statistical analysis of register and sublanguage, these features would be counted and assumed to be similar to usage in the general language.

The fact that fewer pronouns would be used in a chemistry text might be incorrectly interpreted in a statistical count as an absence of referential cohesion (an important feature of Biber's 1996 approach to register analysis). And although I find below that there are significantly more prepositions in the Pharmaceutical Sciences Corpus in relation to the general language (see Appendix 1), it does not follow that the functions of prepositions in general English are replicated in the corpus. In one of the first corpus studies of scientific texts, Sampson and Haigh (1988) found that noun phrases, prepositional phrases, past participles and non-standard *as* clauses are more common in technical writing than in fiction. But it is significant that they argued against characterising these features as 'tell-tale constructions' (1988:218). All of this is of course predicated on the analysis of single, isolated grammatical features or categories. No study has so far been applied to the relative interaction of words between genres, and it seems that there is even more scope for differences between the collocations of scientific English and General English. My preference for Swales' 'genre' therefore reflects a concern for the contextual analysis of certain features, and suggests that even if a feature is equally frequent in two different varieties, its functions and distribution of use are not necessarily the same. This point is taken up again in our discussion of the corpus analysis of grammatical items.

Another reason for adopting the genre analysis approach, is that Swales has established a tradition of analysing research article sections, not just research articles as a whole. Such attention to 'subgenres' has only been tentatively explored in recent corpus work (Biber, Conrad and Reppen 1998). Before the introduction of large corpora, Swales (1990:134) showed that rhetorical sections (Introductions, Methods and so on) have consistent and predictable rhetorical structures of their own. While the model is well known and has in many respects been surpassed by later work (Swales 1998), it remains the first characterisation of science writing that emphasises differences in wording and style rather than the assumption that the text has a consistent system of expression throughout. Swales' work was followed by a number of studies extending his concepts to the entire research article genre and also examining different lexico-grammatical features from the point of view of 'discourse coherence'. In order to give a broad picture of the research article, I summarise some of these studies below, separating those studies which examine the research article as a whole from those which explore specific sections. Since my main method is to analyse the role of collocation from one

section to the next, it is important to set out here a picture of the general linguistic properties of each part of the research article in turn. To avoid confusion subsections of the text (known as rhetorical sections) are henceforth indicated by an initial capital letter: Title - Abstract - Introduction - Methods - Results - Discussion.

3 The Research Article Genre

Swales' work remains the most detailed analysis of the inner workings of the research article genre. In the context of the massive flow of written data in science, Swales sees refereed journals as the 'traffic officers' (1991:94) of scientific information: articles are channelled to the appropriate journals on the basis of how original or significant they are perceived to be in the discourse community. In the case of the research article each specialism has its own conventions regarding graphic and textual format as well as devices for academic accreditation and citation (Swales 1990:6). Despite these differences, Swales claims that there is a fundamental underlying rhetorical system.

At the discourse level, Swales identifies a stereotypical rhetorical structure that is analogous to the knowledge structures of Schank and Abelson's (1977) scripts and Van Dijk and Kintsch's (1989) textual macrostructure. In particular, Swales (1981a, 1990) proposes that the rhetorical structure of Introductions in research articles from a series of different specialisms can be characterised by a macrostructure of one global purpose: to create a research space (the CARS model). This aim is realised in obligatory and optional stages in the argumentation of the text that Swales terms Moves (obligatory) and Steps (optional) (1990:137). Since moves are rhetorical in nature they represent a summary of many different pathways that the argument of a text can go through. The first move, for example, 'establish a territory' is made up of a series of steps which introduce specific areas of the research field as important and relevant to the study, as well as stating the general topic of the study and items of previous literature.

The linguistic features of move 1 include time references to previous research (adjuncts of time such as *recently,* and use of the present perfect), evaluative statements of importance or interest to the field (*it is well-known that*) (1990:144) or, specifically in step 2 statements of amount or quality of evidence established in the field (1990:145). In step 3 the linguistic resources consist of a specification of previous findings followed by a temporal qualification, reporting phrases (*was found to be*) or reporting verbs (*show, demonstrate, suggest*), and bibliographic attribution (1990:149). The

second move, 'establish a niche', involves opening up the existing knowledge structure to weaknesses, either by claiming new factors that expose the old model, or by enhancing the existing model in some way. The linguistic characteristics of move 2 involve references to the negative effects of previous methods with grammatical negatives or conjunctions of adversity (*However, few*) and lexical negatives (*fails to, is inconclusive*) (1990:155). Any weaker or marginal steps are characterised by pointers such as *it is of interest that, a key problem is* (1990:156).

The third move 'occupy the niche' carries the topic on to occupy the gap established in the first two The linguistic features of move 3 involve a lack of reference to previous research, explicit metalinguistic references to the research text (*the present authors, in this paper*) and prevalent use of the present tense (1990:160). By stating the aims of the new research and exploring methods, move 3 takes the rhetorical direction into the 'present' research with increasing explicitness (1990:141). It is noticeable that the Introduction includes many topics that are reformulated in the rest of the research article (especially methods and findings). Since this is also a typical function of Abstracts and Discussion sections, the research article emerges not as a linear text developing its argument from one point to the next, but as a series of more or less detailed recapitulations, differentiated by a change in rhetorical emphasis. We have seen in the previous section that the concept of reconceptualisation and reformulation is a also key issue in the development of terminology.

A number of other linguistic studies have been carried out on the research article as a whole. Some work has been carried out on the distribution of lexical items in research articles (Inman 1978, Love 1993). Most research on IMRD sections has however concentrated on rhetorical move analysis or theme-rheme patterns (Nwogu 1989, Nwogu and Bloor 1991). In a different direction, Atkinson (1992) has traced the historical development of the scientific paper and the evolution of the IMRD sections (the core sections of the research article) from letters to editors in the *Edinburgh Medical Journal*.

Many studies have established that grammatical features (most often verbal tense, voice, or modality) are associated with specific rhetorical functions, such as statements about the use of the passive or authorial comment. Gerbert (1970) for example, analysed 24 verbs in English technical writing, and found that the present represents a limited set of meanings (scientific laws, processes and repeated actions, definitions, descriptions, observations and material properties). The perfect aspect is used to indicate relevance to the research process. Oster (1981) found that non-finite verbs tend to be used for attribution and definition as pre-modifiers (*tumor-derived*

factors in...) or in non-finite clauses (*lipid mobilization in supplying fatty acids*.). Sager et al. (1980:218) found that when non-finites are in end-of-sentence position (typically a clause position reserved for new information), they signal a result (*...leaving all the gears exposed*). Wingard (1981) analyses verb usage in 15 medical texts, showing that up to 40% of verbs occur in the passive, and that while the present indicative is the most frequent verb form (28-40%), 64-78% of verb uses are non-finite (70-80% of which are past participles modifying noun phrases). Hanania and Akhtar (1985) obtain different results from 20 MSc theses showing a preponderant use of the past tense in Methods sections (usually in conjunction with the passive). Malcolm (1987) makes an important distinction between rhetorical constraints on grammar and rhetorical choice. An authors' use of the present tense for generalisations, the past for specific experiments and the present perfect for footnotes are all constraints and unmarked choices. On the other hand, a number of marked choices are available for talking about the work of others. Writers use the simple present or the past in describing previous research as either specific or theoretical, and use the present or the present perfect to distance themselves from previous research (1987:38-40). In addition, Gunawardena (1989) discusses the multi-functionality of tenses such as the 'retrospective' present and the 'inclusive' present. Tenses cannot simply be seen in terms of deictic time reference but also in terms of authorial evaluation of the information he or she is setting out.

The semantics of verbs and the use of modal verbs in 'hedging' have also attracted a considerable amount of research. Thompson and Yiyun (1991) for example classify reporting verbs in research articles, distinguishing between author's stance (where evaluation ranges from praising to negative) and writer's stance (where statements are accepted as fact or non-fact). I. A. Williams (1996) analyses lexical verbs in a corpus of eight texts and establishes differences in phraseology across two types of medical research article. He found that in different rhetorical sections, reporting verbs are more assertive in clinical texts while more tentative in empirical texts. Interestingly, within the context of my previous discussion of 'discourse coherence', he reflects on the differences of lexical choice in different research specialisms:

> [...] the differences in the communicative purpose and its textual realization between medical research types has been much greater than previously assumed [...] (I. A. Williams 1996:195).

In corpus linguistics, research articles have tended to be subsumed in general categories of scientific text (including popularisation). Barnbrook (1996) notes that sublanguages as such have not been analysed in great detail,

largely because scientific texts are treated as whole units and placed together in order to arrive at coverage of several fields (with the assumption that they are all related by degree of specialism).

However, there has been much corpus analysis of research articles in the fields of terminology (Thomas 1993, Pearson 1996) and there is a growing amount of corpus-based discourse analysis. In a corpus analysis of eleven texts on oceanography, Banks (1994b) analyses the distribution of the passive, personal pronouns, modal verbs and lexical hedging (in verbs and adverbs) across rhetorical sections. He finds that there are phraseological differences between modals such as *can* and *may* and that a high proportion (69%) of modalised mental process verbs are used in the passive (*it is believed that...*). He also notes that the lexical hedging of verbs with adverbs (probably, generally) is so widespread towards the latter part of articles (Results and Discussion sections) that their effect is at times redundant. Myers (1989) has argued that such hedging is obligatory when the author expresses some imposition on the community (claims, denials, coining of new terms, apologising for speculation). More recently, Varttala (1999) has compared hedging devices in a 50 text corpus of popular science and technical research articles. All of this evidence of 'hedging' suggests that a conventional voice has become entrenched in science writing, a point that is supported by work on collocations and phraseology.

Corpus analysis on lexical collocation in research articles has also been undertaken, either taking a phraseological perspective or concentrating on typical NP complements of verbs. Zambrano (1987) analyses the phraseological patterns common to Abstracts and Discussion sections, including phrases identifying general problems, concerns of the research article (*this article / paper / study* etc. *shows /suggests / investigates* etc.), findings (involving nominal comparatives with *show*) and implications (involving a high degree of modality: *the possibility that, the fact that*). Master (1991) finds that inanimate nouns (*shuttle, particle*) are more likely to be the subjects of active verbs than passives, and such verbs are more likely to be verbs of causal processes (*cause, affect, prevent*) than reporting verbs (*show, indicate, suggest*) (a distinction echoed in the PSC - the research article corpus, as described later). Other work concentrates on the clause patterns associated with certain families of nouns (Dubois 1981, Francis and Kramer-Dahl 1991).

A small number of studies address the use of grammatical items and cohesive devices. Thyman (1981) proposes that the description of non-linear (simultaneous) events in scientific writing has led to changes in the use of specific cohesive devices, such as the classifying and defining function of *this. This* is widely used in the process of reformulation, a point noted in the

corpus study below. From a more phraseological perspective, Abraham (1991) distinguishes between the use of *because of*, signalling given information, and *because* (a signal of new information). *Because of* is the preferred expression in scientific writing (41% of the occurrences) as opposed to 6% in spoken discourse, suggesting that reformulation of given data is an important function of scientific texts.

Biber, Conrad and Reppen (1998) carried out a cluster analysis of grammatical features on a corpus of 20 different scientific research articles. Using Biber's (1993) concept of multidimensional analysis, Biber et al. (1998:157) demonstrate that ecology articles have relatively more impersonal features (conjuncts, agentless passives, past participle post-nominal clauses and adverbial subordinators) and more narrative features (past tense verbs, synthetic negation, present participle clauses) than a similar corpus of research articles in history and a corpus of general fiction. When different rhetorical sections in their corpus are analysed on the Impersonal / Non-impersonal scale, they find perhaps suprisingly that Discussions are the most impersonal, followed by Methods, Results and Introductions. Their explanation (that Discussions frame other researchers' work in the passive: 1998:168) is interesting, although multidimensional analysis places much emphasis on features of science writing that are well-documented in the literature (passive verbs, tense, past participles). There seems to be little scope in their work for the analysis of less salient features such as hedging (the use of modals) or *to*-clauses in science writing, as these are characterised in their statistical analysis as typical of other registers. Nevertheless, this is the first parallel analysis of a battery of linguistic features within the research article genre. Biber et al.'s (1998) study underlines the fact that much work on research articles as a whole has concentrated on the linguistic features of verbs, the overwhelming majority dealing with tense and voice (the passive). This is perhaps not surprising, in that tense and verb form are key elements in signalling the attitudes of the author.

Many other aspects of scientific discourse have been carried out in the context of specific rhetorical sections. A brief survey of each rhetorical section is set out below.

3.1 Titles

Very few studies have concentrated on research article Titles in their own right. Apart from observations of their highly condensed nominal style, little is known about the relationship between the Title and the rest of the research article. Generally speaking, Titles are seen as sources for keywords in the

information sciences. For example, Diodato (1982) has studied the relative frequency of Title words in 50 chemistry, history, mathematics and philosophy papers. Her findings indicate that 70-80% of all Title words occur in the Abstracts and the first paragraphs of articles. She finds that chemistry papers are the only papers to have an increase in the amount of Title words throughout the paper, with the largest increase in the final reference sections. The implication is that Titles are a good indicator of subject-matter, but Diodato has little to say about the role of the Title in staking out the research article's claims.

In a rare analysis of research article Titles as a subgenre, Jaime-Sisó (1993) examines a corpus of 2 000 journal Titles from six fields of medicine (all downloaded from the electronic indexing service MEDLINE). Jaime-Sisó is particularly interested in grammatical change over time. She finds that from 1980 to 1990 the number of Titles with active clauses (e.g. *Dietary fish oil delays puberty in female rats*) rose from steadily 0% to 40%. She observes that these Titles are used in dynamic areas of science (developmental biology) and in high prestige journals with consistently high scores on the impact factor scale (Williams 1996, see section 3 below for an explanation of 'impact factors'). Jaime-Sisó also finds that the types of verbs involved in these active-clauses (*contribute to, is required for, contains*) do not give empirical facts or findings as such, but oblige the author to justify the novel results elsewhere in the article. The Title effectively becomes a promissory notice of results. The point here is that linguistic change reflects the changing role of the Title in terms of its environment. Titles have to 'compete' for readers' attention, and the use of Titles to suggest (if not carry) significant results corresponds to the growing use of graphic abstracts in chemistry and in other fields. This also implies the increasing independence of the Title and Abstract as 'stand-alone' text types, a concept introduced by Gläser (1991). Jaime-Sisó is careful to note that the occurrence of active verbs has only become prevalent in a restricted field: other fields have significantly not been affected by the trend. These observations require more extensive comparative work, but do provide an interesting picture of the Title as a key element in the framing of scientific claims. Although Titles do not normally set out a propositional argumentation as such (unless they contain a full clause, as Jaime-Sisó has demonstrated), they clearly have a function in situating the research article in a wider framework and one might assume that Titles vary in ambition, from setting out very specific technical points to evoking or questioning the general *status quo*.

3.2 Abstracts

The Abstract is considered to be one of the most important sections in the research article genre. The Abstract represents the main ideas of the text, and is often seen as an independent text in its own right. Abstracts are routinely reproduced without the main article in abstracting indexes. As a result, more research on Abstracts has been undertaken than on other sections, largely in the information sciences and in fields such as textlinguistics. Most linguistic studies find that Abstracts are highly polished and condensed texts, with a high frequency of relative clauses and nominal embedding which makes them particularly difficult for non-specialists to read. Not surprisingly, Abstracts are seen as prototypical scientific texts, a fact that may artificially obscure the role of those sections of the research article which tend to be more accessible (Introduction and Discussion sections).

Most work centres around the processes involved in summarisation, and tends to concentrate on Abstracts produced by a third party (either professional abstractors or students). Baker et al. (1980) have analysed the role of professional abstractors at the *Chemical Abstracting Service* (CAS). The abstracting business is said to be immense: CAS alone employs over 2 000 indexers (Metanomski: personal communication). The size of the business is reflected in the number of guidelines designed for abstractors (Weil et al. 1963, Borko and Chatman 1963, Cleveland and Cleveland 1983, Cremmins 1982 and Memet 1986). Khurshid (1979), Polskaya (1986) and Raya (1986) have all examined indexing abstracts from the viewpoint of information science, usually examining the most successful strategies for creating informative abstracts. Typical of this kind of study, Buxton and Meadows (1978) set out the common points of information contained in chemistry Abstracts. Rush et al. (1971), Pollock and Zamora (1975) and Sharp (1989) also discuss the possibility of producing automatic abstracts. Automatic abstracting has been influenced by Van Dijk and Kintsch's (1993) propositional textgrammar and de Beaugrande and Dressler's (1981) studies on summaries formed by the matching of textual patterns. Gopnik (1972) set out an exhaustive textgrammar of technical Abstracts from this perspective. She sets out propositional 'macro-rules' which resemble Swales' (1990) rhetorical moves and steps.

Much linguistic work on Abstracts concentrates on the quality of summaries produced by students (Frank 1971, Fløttum 1985, Sherrard 1989). Meyes (1990) find that non-expert summarisers delete the wrong information and construct propositions on false premises because they lack background

knowledge of a specialist field. Gibson (1992) and Drury (1991) have both demonstrated that non-author Abstracts which are perceived to be successful tend to have topical sentence themes as opposed to textual and interpersonal themes. Drury (1991) finds that rather than simplifying texts, summarisers tend to render themes more abstract and technical (1991:436). The successful summariser also reduces the number of relational and embedded material verbs from the original text, introducing more material processes at the rank of clause (1991:447: i.e. from *It is thought that the temperature rises* to *The increased temperature...*). This is mirrored by increasing lexical density and use of grammatical metaphor in successful summaries (Drury 1991:448). Similarly, Salager-Meyer (1990b) finds that unsuccessful Abstracts are particularly difficult to read, partly because they omit important moves (*conclusions* or *purpose*) or order them in unexpected ways (*results* before *purpose*, *conclusion* before *results*) and partly because the 'valuable signposts' of discourse signalling and cohesive devices are usually absent in Abstracts (1990b:378).

There has also been much descriptive linguistic work on a typology of Abstracts. Generally, two main forms are recognised. The informative Abstract introduces the main ideas and explains the essential points of the original article. The indicative Abstract on the other hand reformulates the article, following the progression of the article as closely as possible. Informative Abstracts in particular are said to use markedly different expressions and terms than the original text (Cleveland and Cleveland 1983:4). Grätz (1985) claims that most Abstracts in the sciences follow the rhetorical structure of the original text closely and serve as indicative Abstracts. However, Gläser (1991) has argued that the Abstract is a separate genre rather than a rhetorical section, and points to its condensed presentation of content and lack of deictic reference or stylistic devices. Endres-Niggemeyer (1985) suggests that authors do not follow journals' instructions on Abstract and IMRD sections in any case. She argues that the categories suggested by journals do not cater for the needs of the reader, and that authors tend to structure Abstracts and other sections according to their own specific objectives. This is an interesting observation, suggesting that rhetorical sections are less clear cut than Swales and others have assumed, and that scientists impose their own rhetorical goals rather more freely than might have expected. Endres-Niggemeyer proposes conceptual text types situated around topical poles, such as the *overview* and *model building* Abstract versus the *practice oriented* and *theory-descriptive* Abstract (1985:45). These are the modes of discourse successfully adopted by authors rather the kinds of text requested by journals.

Descriptive studies of Abstracts have also compared the linguistic features of different types of Abstracts, and a smaller number have compared the Abstract with the rest of the text. Bernier (1985) and Craven (1965) have set out the syntactic features of what they call the 'terse literature'. Harris (1985) examines authorial comment and stance in scientific Abstracts, and Sastri (1968) analyses prepositions in chemical Abstracts. King (1976) sets out the typical vocabulary profile of author Abstracts. Dronberger and Kronitz (1975) and Reder and Anderson (1980) studied the readability of indexing-abstracts as a function of vocabulary. In a rare piece of comparative work Fidel (1986) analysed vocabulary differences between indexing-abstracts and Discussion sections of the original article. In an similar comparative study, Nwogu (1989) analysed cohesion, thematic progression and Swales' system of moves in 15 medical research articles, compared with their Abstracts and popularised journalistic versions. He finds that Abstracts have two obligatory moves (*indicating consistent observations / stating research conclusions*) and seven optional moves (corresponding to Salager-Meyer's moves of *purpose* and *methods*: *presenting background information / reviewing related research / describing data-collection / describing experimental procedure / highlighting overall research outcomes / explaining specific research outcomes*) (1989:171). Abstracts do not include the moves *describing the data-analysis procedure* and *indicating non-consistent outcomes* (1989:161). Nwogu also finds that Abstracts have a much lower density of sentences per move (2.02) compared to research articles (4 sentences/move) which is reflected in the complex clause structures and a greater sense of embedding or 'compaction' in the Abstract (1989:180).

In a computer-based analysis of technical Abstracts, Kretzenbacher (1990) examines a corpus of 20 Abstracts with their original academic research articles in German (a total corpus of 88 000 words). He confirms the general finding that Abstracts have a highly nominal style, with a significantly higher noun-per-sentence ratio, more 'verbal substantives' in German (which are usually marked by the equivalent of noun suffixes *-ness*, *-ity* etc. in English), and more nominal compounds than the original article (1990:56-67). The main articles are found to have a significantly higher range of finite verbs, while Abstracts have relatively more passive forms. Interestingly, Abstracts tended to use as many modal verbs as the main articles. Only 8 of the 20 articles were found to have more modal verbs than their Abstracts, a finding that suggests an affinity between with Discussion sections, where results are frequently summarises, reformulated and re-presented. Abstracts are found to have a slightly lower word per sentence ratio than the main texts, (23.8 to 24.62) which is still high in comparison with other German genres (1990:86), presumably because Abstracts make relatively more use of embedded clauses

rather than longer clause complexes. Also, Kretzenbacher finds that Abstracts tend to use nominal groups and finite verbs as attributive elements of clauses, a typical construction in German (1990:101). Kretzenbacher also finds that Abstracts have relatively more genitive attributes (part of the general nominal style in German) and definite articles, while the main texts have relatively more infinitives, anaphoric reference, and personal deictic reference.

In the first of a series of large corpus-based analyses of Abstracts, Salager-Meyer (1992) analyses verb tense and voice usage and modality in 84 Abstracts (from 49 research papers, 21 reviews and 14 case reports). She finds that the active past tense is the most frequent verb form (51% across all types) and corresponds with the rhetorical moves of *purpose, results, methods* and *case presentation*. The past passive is particularly prevalent in the *methods* move, indicating that this is an obligatory form of expression. In the *purpose* and *conclusion* moves on the other hand, Salager-Mayer finds that the choice of tense is more open to rhetorical interpretation: the present may be used to state basic truths, but also to emphasise that previous research is relevant to the study. The present perfect also has a multiple function of *reference to past experiments*, *introducing a topic* as well as *distancing the author from the findings* (1992:106). The past tense is found to be much less prevalent in moves of *statement of the problem* and *data synthesis*, where the function of the past is to indicate the undeveloped nature of previous findings. Finally, modality is also found to be move-related, with the most frequent modal, *may*, indicating a high probability of claims in the conclusion; *can* being associated with data synthesis, and *should* used in preference to other modals in the *recommendation* move (1992:105). Such a consistent use of verbs for rhetorical purposes (in tense or modal form) further supports Swales' observations about the controlled nature of scientific discourse, but also suggests that tenses and verb forms imply a much more sophisticated set of interpretations than was previously thought.

3.3 Introduction Sections

The Introduction section has been a privileged area of linguistic analysis since the early work of Swales (1981a). Yet Introductions are sometimes seen as redundant parts of the research article, since specialists claim that they tend to skip them. Ironically, the interest in research article Introductions therefore lies in the fact that they appear have a primarily rhetorical purpose, often linked with the need to provide academic validity to the article as well as a useful background for readers who are non-specialists (Kinay et al. 1983). For a variety of reasons, therefore, Introductions are seen as having a

relatively freer style than other research article sections and are also considered to provide the writer with a certain degree of stylistic freedom.

Apart from Swales' (1990) analysis of Introductions set out above, West (1980) has studied the use of *that*-nominals which are relatively more frequent in the Introduction section as opposed to the other rhetorical sections. Hanania and Akhtar (1985) found the present to be the usual tense in the Introduction, associated with the functions of introducing background, establishing assumptions and the purpose of the research. Gunawardena's (1989) analysis of 10 biology and biochemistry articles shows that the present perfect is particularly prevalent in Introduction and Discussion sections, where both sections relate shared experience as well as report past research. In their analysis of 15 medical research articles, Nwogu and Bloor (1991) found that Introduction and Discussion sections have overlapping thematic structures (associated with explanation and argumentation) while Methods and Results sections have relatively constantly changing theme structures (associated with description). Finally, the similarity between Introduction and Discussion sections has been often noted, especially in terms of phraseology and use of modal verbs (Salager-Meyer 1992, Williams 1996, Gledhill 1996).

3.4 Methods and Results Sections

Methods and Results sections are the most inaccessible parts of the research article to the non-specialist. However, for the expert reader these sections usually constitute the first port of call, especially in the experimental sciences. While few studies have concentrated on these sections in their own right, a small number of comparative analyses have been carried out. Generally speaking, Methods sections are found to be predictable and repetitive, and generally set out procedures as well as detailed findings. It is well known that Methods account for the vast majority of passive verbs, especially in chemistry (Hania and Akhtar 1985). Ironically, findings are not always fully set out in Results sections, which are generally limited to reformulating the Methods and summarising quantitative observations and statistics. Evaluation and interpretation are reserved instead for the Discussion section. Practices vary considerably from one journal to the next, and sometimes these sections are combined or accompanied by supplementary sections known as 'Materials and Methods', 'Experimental' or 'Results/Discussion'.

For Swales (1990), Methods sections constitute the core science of the research article. In most cases, especially in structural chemistry, the

Methods section is the linear version of the laboratory book, a listing of procedural formulae with details of techniques, brand names involved, temperatures, measures, amounts used, reaction speed, molecular size (mml, mhz, mmo) and so on. Swales claims that these sections are 'highly abstracted reformulations of final outcomes in which an enormous amount is taken for granted' (1990:121). Swales points out that this seems to belie the empirical ideal in which massive detail ensures the possibility of replication. The Methods section carefully legitimises the rest of the article, and in Swales' view constitutes a rhetorical section just as much as any other. More generally, the passive is commonly said to enable a distancing of responsibility of actions from the actual protagonists, as we discuss later in terms of grammatical metaphor (Sager et al. 1980:209, Swales 1990:120).

Few studies of Results are conducted without reference to other sections, and according to Swales both Methods and Results sections are 'mutually inter-dependent' (1990:121). The literature usually points to linguistic similarities between both. Adams-Smith (1984) analyses authorial comment (in terms of modality items such as *possible*, first person references, markers of analogy such as *like, similar*) and finds that the distribution of these items throughout IMRD sections decreases in the Methods and Results sections and increases again in the Discussion section. She also finds that past and passive verb forms follow this pattern, and her results on the distribution of the passive in Methods / Experimental sections are echoed by Banks (1998). West (1980) has also demonstrated that *that*-nominalisation is extremely rare in Methods and Results sections, while relatively frequent in Introduction and Discussion sections. This is corroborated by Brett (1994) in his analysis of Results sections in geography research articles. Finally, Heslot (1982) and Wingard (1981) have shown that the simple present tense is more frequent in Introduction and Discussion sections, and the simple past tense more frequent in Methods and Results sections. The other complex tenses (continuous / progressive) are rare. According to most of these studies, Methods and Results sections tend to be conceived as the most 'scientific' sections of the research article, i.e. the most removed from general prose and other varieties. However, Biber et al.'s (1998) observations of relatively high amounts of Impersonal features in Discussion and Methods sections (with Discussions scoring very highly on the Impersonal scale) serves as a warning not to take single features as indicative of absolute similarity between two sections. It is possible that superficial similarities (especially in verb form) do not correspond to deeper differences in rhetorical structure: Results sections deal with the same themes as Methods, but set them out in fundamentally different ways. Some of these differences may become clearer in our discussion of collocation in section III.

3.5 Discussion Sections

There have been a number of studies of Discussion sections (McKinlay 1983, Hopkins and Dudley-Evans 1988), largely from the point of view of rhetorical structure. Some comparative studies have emphasised the similarity of grammatical features with Introduction sections (Gnutzmann and Oldenburg 1992). On the basis of a 20-text corpus, Dubois (1997) examines a typology of clauses (establishing semantic categories such as *metatext, methodology, conclusion, comment*), rhetorical move analysis and hedging. She argues that the rhetorical functions of Discussion sections are very different to Introductions, since the Discussions provide a detailed synthesis of results and their evaluation as viable elements of a new model. Swales (1990) suggested that Discussion sections are the mirror images of Introduction sections, looking out from the research into the wider world. Thus Introductions synthesise past research and evaluate old models inwards towards the 'core' scientific activity (Methods-Results), while the Discussion section does the reverse, returning the product of scientific research to the discourse community. This does not explain why grammatical features are shared, although as with Methods and Results sections, we have seen that superficial similarities of single grammatical forms are not always indicative of deeper rhetorical differences.

4. The Discourse Community

In the previous sections, I have set out an introduction to the theory of the terminology and discourse of science. In this section I examine these theories in the context of a cancer research laboratory. In the first part, I explain the context of cancer research and set out a basic explanation of cancer with a view to defining the discourse of cancer research itself. I then conduct a survey of cancer researchers, designed in part to provide a context for the corpus set out in sections III and IV. Given that many of my informants have themselves contributed their texts to the corpus, any light they can shed on the writing process and their use of research articles is relevant to this study.

[From this point, in order to differentiate their opinions, researchers are referred to by their italicised initials (as listed in the preface). Research papers have been given a code indicating which journal they come from (e.g. TL, BMJ5, CAR1), with a number when there is more than one article from

that journal. These correspond to the titles and bibliographic data listed in Gledhill (1995b) and in Appendix 2].

4.1 The Discourse of Cancer Research

A major linguistic motivation for studying pharmaceutical and cancer research is that these fields involve a high degree of abstract pharmaceutical knowledge. The interaction between a knowledge structure and the language in which it is couched is of particular interest to the phraseologist. In this section therefore I attempt to establish the discourse of cancer from the point of view of the scientists themselves. This is a one-sided view of discourse, in that it is seen as engendered by scientists for scientists (with no participation with patients, or public bodies, for example).

Cancer research is perhaps one of the best funded and most influential research activities in medicine. The nature and reputation of the disease is emotive and dramatic, and this is reflected in the large amount of charity fund raising and publicity that is generated for medical research in this area. A review of the Science Citation Index (SCI 1993) reveals that cancer research is the most important single specialist topic in medicinal research. The SCI lists journals in terms of their importance, largely measured by citations and cross-citations in other periodicals. The SCI lists over 8000 journals, and medicinal applications of biochemistry account for two thirds of the first 100 on the list. Of the first 600 journals on the SCI list, 18 (3%) have cancer or oncology in their title. Other diseases on the other hand have on average only one journal-specific title in this list (two for AIDS, one each for Arthritis and Rheumatism, Heart disease, Leprosy, Schizophrenia, *inter alia*). Thus medical science is one of the biggest areas of scientific research, and cancer research in turn can be seen to be one of medicine's most prominent activities, at least according to the 1993 listing. Cancer research appears to be an enormous research programme, and the amount of money invested in the disease, at least in the West, reflects an increased awareness of the effects of cancer on an aging population. As noted by Kevles (1995), in the same way that space exploration was given an artificial boost in America in the 1950s and 1960s, cancer was not a major area of medical research until it enjoyed political backing during the 1970s in Nixon's 'War on Cancer'. Cancer is therefore at the centre of global scientific activity, and the discourse of cancer is very highly politicised.

Most cancer researchers agree that the problem with the public perception of cancer is that it is not one but many diseases. Cancer research covers a broad sweep of specialisms (drug synthesis, virology, biochemistry,

population genetics, patient care etc.). Various research activities (chemotherapy, metabolism studies, causal nutrition studies) contribute to solutions leading to the ultimate medical goal: the cure for cancer. Other researchers, by the journals they read and publish in, tend towards the description of the problem (such as oncogenesis, cancer epidemiology and virology) while others look at the side-effects and long-term issues associated with the treatment of cancer (toxicology, palliative care). This complexity poses an obvious problem for terminologists, and also explains how difficult it is to consider the Pharmaceutical Sciences Department as a clearly-defined discourse community. In Swales' terms (1990:32), the discourse community is fragmented and has differentiated goals. In terminologists' terms, cancer is a distributed concept, occupying a series of relative positions rather than a central role in and of itself. From my survey of the Pharmaceutical Sciences Department , two defining features of the discourse community emerged:

1) Scientists situate themselves in a network of professional relationships.

The extent to which individual researchers associate themselves with cancer research or chemistry is a complex issue. The chemists in my survey explained their approach to the problem in terms of combating disease with target drugs, growth inhibitors and antiviral agents, while the molecular biologists talked in terms of finding new approaches to the disease by understanding such processes as cell death, replication and differentiation. Since cancer researchers often commission structural analyses from chemists, the two research programmes can be seen to be systematically interrelated and one might establish from the beginning a professional 'service' relationship where the oncologists (working *in vivo*) require functional and structural analyses of pharmaceutical substances from the chemists (working *in vitro*).

2) Scientists situate their research in a rhetorical relationship to cancer.

The idea that there are some researchers 'close to' cancer research with others at the periphery is only a partial picture. In the survey this became an question of how the researchers justified themselves to an outsider. In my survey, only five informants declared themselves cancer researchers. It emerges therefore that a community of cancer researchers can not be defined by institutional or social arrangements alone, and that it necessary to refer to a notion of the scientific model of the disease itself. In order to give an insight into how the core phraseology of cancer research is formulated, I have set out below an introduction to the science behind the disease. The text reveals the dialectic which exists in the wider research community regarding the nature of cancer as a medical and scientific

problem. The text is based on my discussions with expert informants (most notably, *MT*) and on an influential recent introduction to the subject by Thomas and Waxman (1995). The key terms which typically occur in the corpus have also been italicised:

The science of cancer.

All cancers have in common a genetic virus. This is promulgated by a potentially malignant part of a gene: the oncogene. The virus produces defects in the ways cells are reproduced and developed according to their predetermined function in the *metabolism* (the un-diseased process being termed *differentiation*). Cancer is the physical effect (by *proliferation* or *tumour growth*) of a breakdown in this genetic process (*carcinogenesis*) and in particular the *overexpression* of the oncogene. The cause of *malignancy* in the oncogene can take place at any place within the cell or in its immediate environment. This complexity accounts for a wide variety of specialist research, going beyond the field of genetics and involving the organic chemistry of compounds that come into contact with the cell. For example, malignancy involves *growth factors* attaching themselves to the surface of the cell, and also the *activation* of oncogenes in the cell nucleus where '*ras*' *proteins* are able to transform DNA within the nucleus.

Above the level of the cell, the causes of these changes become less identifiable as the physiological system becomes more complex. For exemple, genetic changes have been known to be caused in breast cancer by steroids and *peptide* growth factors. These are complex chemical proteins such as *kinases*, often described as a cloud of toxicity. There is however no consensus on the molecular origin of malignancy (Thomas and Waxman 1995: 6). The only generalisation appears to be that diet is by far the largest cause of growth factor activity, followed by tobacco consumption, viral infection and environmental influences (such as electronic radiation). Recently, debate over the causes of cancer has been hampered by empirical problems. Although many human tumours are known to be caused by DNA-related viruses (for example, *immunodeficiency* virus is associated with AIDS-related tumours), most scientific research has concentrated on simpler animal *RNA* viruses (1991:5).

Because of the uncertain nature of malignancy, *pharmaceutical responses* to cancer are varied. Generally, *intervention* in the genetic processes is not regarded as viable (1991:14), since genetic breakdown is activated by external factors. Instead, it is the actual moment of *activation* and the consequent production of cancerous genes (*expression*) that is the target of pharmaceutical cancer research. There has generally been particular emphasis on the study of processes just on the surface of the cell, where *growth factors* interact with a cell's chemical *receptors*. Other researchers are interested in the transfer of chemical information achieved by chemical *synthesis*. Yet another group of researchers are interested in the possible *starvation* of the tumor's own *metabolic system*. By developing compounds that can *target cells* and replace *receptors* or *growth factors*, a receptor can be developed that destroys the incoming *growth factor* by *inhibition*

> (a *tumor necrosis factor* , for example destroys *carcinogenic receptors*). Given that there are over 2 million receptors on one cell, there is considerable scope for specialism in different types of inhibitors.

This is the everyday language of cancer research. By introducing the central terminology of cancer in this way, it is possible to build up a knowledge structure of the field. It is also of no surprise to find much of the basic phraseology of this text within our text corpus, especially in Introductions and explanatory sections of the research articles. Such an account explains why seemingly innocuous semi-technical expressions such as *activation, expression, inhibition* appear to be involved in much of the recurrent phraseology in the corpus.

The knowledge structure of cancer appears to be oriented into two semantic planes. Firstly, research can be situated as a spatial metaphor to the parts of a cell the researcher is most concerned with, such as the molecular processes within and surrounding the cell. Secondly, research can centre on the description of the effects of the disease, or causality and chemical intervention against the disease. Research can thus be entity-oriented (around the object of the *cell*) or event-oriented (around the chemical processes and wider effects of the disease). For example, many of researchers in the PSD were concerned with inhibition at the surface of the cell, and this view of the disease may not correspond exactly to other researchers in other departments or institutes. As a consequence, our text corpus tends to cover a much broader range of issues than are of current concern to the PSD researchers, although it can also be seen to represent a reasonable range of research questions that have been formulated about the disease in general.

Given the scope and the immense activity involved in cancer studies, it is easy to see how scientists need to be very specialised in order to claim any expertise or centrality in their own particular field. The Pharmaceutical Sciences Department can therefore only represent one tiny fragment of a larger research programme. In this context, it would be useful to discuss the dynamics of the discourse of cancer research, and in particular the ways individuals and groups of researchers gain attention and claim relevance in such a vast discipline.

4.2 A Textography of the Pharmaceutical Sciences Department (PSD)

This section describes some of the problems encountered when one considers the extent to which a corpus can be ‘based on’ a very specific discourse community.

Language in Performance Series No. 22, Tübingen, Gunter Narr Verlag, 270pp.

The Language Studies Unit at Aston University, where I was based, is situated conveniently near the Pharmaceutical Sciences Department. It was this connection that led me to contact the PSD with an initial questionnaire about how the scientists used language professionally. The fact that researchers in the pharmaceutical sciences were easily accessible and interested in the role of language in their work was a considerable advantage in building the corpus. The researchers gave free access to written research and publicity material, including departmental listings and press cuttings. They were also happy to talk about their texts and their use of language, and to see that their activities aroused interest in other parts of the university. The ethos of the discourse community is, I believe, an important methodological step in building very specialised text corpora. This has also recently been a key feature of the approach advocated by Swales (1998), in which a textography is based on dialogue and mutual exchange of ideas in order to better understand the constraints on the production of texts and the context of use of specific text types.

None of the PSD had time to undertake more than one formal interview (usually lasting one hour); so I decided to survey as many researchers as possible in order to get a broad view of research. The survey is therefore very different to the very close longitudinal study of the type undertaken by Myers (1990). Even though the fourteen people interviewed included only a third of the academic staff in the PSD, the research activities of the department can be considered to be reasonably covered.

The main fields of expertise in the Pharmaceutical Sciences Department involve medicinal applications of chemistry to a number of major diseases (including rheumatism, AIDS and tuberculosis). However, the largest group in the department is the Cancer Research Group, which maintains its own identity. At the time my survey was carried out, the PSD had a large output of research with a number of high profile breakthroughs in the press. According to its promotional literature, the department is working towards 'advances in the understanding of disease in the metabolism' (the sum of all the chemical reactions in the living cell and hence the organism) and 'targeting of disease by the development of highly specialised synthetic compounds' (the artificial production of organically functional drugs). This conceptual difference is represented in an institutional division between departmental sections. In 1992 the size of these groups (not including postdoctoral workers and technicians) was as follows:

Section I: Drug Development
(Pharmaceutical Sciences Institute: 13 academic staff, 6 in the survey).

Section II: Cancer Research, Toxicology and Microbiology.
(19 academic staff and 8 in the survey).

Section III: Pharmacology
(5 academic staff 1 in the survey).

This raises the potential distinction between the discourse community and a community thrown together from the point of view of an institution (a difficulty discussed in Swales 1998). Generally speaking, institutional communities do not necessarily correspond to the notion of discourse communities (defined by 'what they talk about' and social networking rather than by socio-economic grouping). An extensive survey of 20 000 academics by Boyer (1994) has suggested that many researchers in British universities have a greater sense of identification with their discipline than with their own institution. As we have seen above, simply because a researcher is working on a 'cure for cancer' does not mean that he or she defines their own specialism as 'cancer research'. The survey reveals below that the research goals of my informants were not fixed to cancer research *per se* and that researchers did not always respond to the question 'are you working on cancer research?'.

For example, the structural chemists (SF, BF, JG) had recently won a substantial grant from the Cancer Research Campaign - yet during the survey they distanced themselves from cancer research *per se*. Such issues as funding or research group membership is therefore not a clear guide to an individual or group's perception of community, at least as they present themselves to outsiders. To complicate things further, one informant admitted that there was an unofficial policy of understating involvement in cancer research because of potential animal rights protests. In another example, the pharmacist *WF* felt obliged to switch his research to DNA molecules from his more original work on a specific inhibitor because of departmental policy. Did *WF* feel he belonged to the community of 'cancer researchers'? His answer to this was not clear-cut. Such institutional matters of policy and presentation presumably constitute an area of tension in the department, and suggest that a corpus of texts on 'cancer research' is not a truly accurate description of the kind of texts and genres that the scientists see as valid and central to their professional work.

It might be possible to determine which texts to include in a specialised corpus by referring to statistical measures of importance or centrality, such as the impact factor. Such a measure would presumably separate the choice of texts from the personal and subjective feelings of the researchers. As mentioned above, the impact factor (IF) in the Science Citation Index is a statistical measure of the number of references that have been made to a

single research article or journal in a general sample of the literature (sometimes many thousands of journals). It is significant that both individual scientists and research journals are increasingly judged on their impact factor scores. In a survey of IF scores, Williams (1996) found that these scores are often taken into account when evaluating a person's research activity and departmental funding. The system is self-perpetuating in that journals which score highly on the SCI league table consequently attract more research article submissions and, in return, receive higher IF scores. This in turn influences the need to produce persuasive and well edited research articles. While I have used IFs to justify the inclusion of some papers in my corpus, they are not necessarily as reliable and as objective as they seem. As reported below, some researchers were sceptical about the accuracy and relative value of citations as measure of successful research, and had alternative ways of assigning importance and prestige to specific journals and research articles.

In an environment where pharmacists and others are competing for research funding from cancer research organisations at the same time as cancer researchers 'proper', the perceived relevance of a specialism must have a consequential effect on a researcher's place in the hierarchy of his or her field. It is noticeable in the corpus that Abstracts and Introductions often mention cancer research as relevant applications, even when the main focus of the text is on a relatively distant topic, such as crystal structure in inorganic chemistry. The issues of field-centrality and representativeness are discussed later in section III (corpus design).

4.3 Details of the Survey

A questionnaire was prepared and interviews arranged with fourteen researchers from the Pharmaceutical Sciences department. The aim was to gather information on two main areas: the discourse community (4 questions) and the use of texts in that community (6 questions).

Survey question 1). What is your title and position within the Pharmaceutical Sciences department? The survey involves a wide range of scientists: the chief academic administrator (*PRL*), three professors (*MT , WI* and *AG*), two senior tutors (*RL, KW*), one senior lecturer (*PL*), five lecturers (*DP,WF, JG, SF YW*) and three research fellows (*DA, HM, RW*).

Survey question 2). What is your specialism, the main field to which you would say you belong?

The symmetrical way the scientists fit into the department's research groups was not echoed by researchers' opinions about their own specialism. All the members of the Cancer Research Group described themselves first as microbiologists, and stated that their general expertise was in cancer research (*MT, KW, YW* metabolic effects of cancer, *PL* cellular properties of tumours compared to other diseases, *AG* chemotherapy and cellular delivery of drugs). Another three microbiologists were interested in cancer and how its treatment affected their own discipline, citing expertise in enzymology (*PRL*), cell differentiation (*DP*) and developmental biology (*RL*). On the other hand, the pharmacists and chemists also cited cancer as the first of many applications of the synthetic molecules they are designing. *WF* is an expert on the synthetic production of organic compounds that are part of the chain structure of DNA, as well as cyclic compounds that can inhibit carcinogenic factors. *SF, WI* and *RW* are each interested in the link between growth inhibition and a specific family of compounds (phosphates). *JG* is concerned with the synthesis that takes place between medical compounds and their target sites. *DA* is interested in the structural elaboration of chemical chains, with long term medical applications.

The perceptions of researchers about each other also made this a complex issue, *RW* describing the 'pure chemist' *WF* as a cancer researcher. As noted above, these differing perceptions arise from the complexity of the problem, and from the seeming impossibility, within the field, of conceiving of cancer as a unitary entity or process.

Survey question 3) How would you describe your field of research in terms of a) its aims?, b) its main concepts or objects of research?, c) its methods?

This question specifically aimed at eliciting 'the common purpose', a central concept of Swales' (1990) definition of discourse community. The microbiologists and pharmacists divided neatly into two groups on this. The cancer researchers and microbiologists stated in general terms the desire for 'better understanding' of disease, involving the complex mechanisms of biochemistry above and below the level of the cell. For example, *YW* stated that the aim of chemotherapy is to find the most effective killer of tumour cells at the same time as the most efficient targeting drug to avoid further damage. Similarly *PL* and *RL* stated that the aim of their research was to understand how intra-cellular mechanisms involving control genes allow for cell targeting. The pharmacists had much more specific aims which required complex justifications, involving a description of specific phenomena rather than an understanding of the whole system. While

they were keen to mention possible applications and diseases, their methods differed more distinctly from their aims than those of the other research groups.

The survey question suggests that informants state the aims and methodology of the research discipline. However, it is hard to see how these cannot also include claims of centrality and individual originality, and this is how most respondents answered it. The phrasing of most of the methods (items such as new, novel, development, accurately) and some of the aims (*WF, MT*) emphasise at least some implicit claim of individual originality within the context of an established research paradigm.

Survey question 4) How does your own specialism relate to those of your colleagues inside and outside the university?

This raises the distinction between an institutional community versus a wider discourse community ('a discipline') and attempts also to establish the 'common mechanisms of interaction' said to define the discourse community. Generally speaking, the scientists constitute much more of a discourse community within the institution than their equivalents would do in the social sciences or the humanities (both areas where research is often perceived as individual activity).

There were clear areas where researchers claimed they worked very closely, and all of these were linked to the production of written genres. Most importantly, all researchers were involved with joint publications (not necessarily within the same research group). Much research in chemistry is published in series (*SF*'s contribution to the corpus is 'Part 7' of his findings) and any joint series of publications must contribute significantly to a sense of long-term common purpose. Most researchers also co-operated on official policy documents within the department which ultimately determined which research group they were working in.

Outside the university, research appears to be conducted in loose groupings, very often of an institutional nature (compare this with generative or functionalist schools in linguistics, for example). *AG* noted that researchers would be aware of related groups elsewhere which would be regarded as 'soft competitors' exchanging research papers and communications, coordinating some grant proposals, at other times competing for them. *WA* stated that for cancer research there were national and international work groups that exchange results and negotiate areas of specialism in order to avoid duplication. *MT* also noted that if exciting laboratory results occurred, colleagues would telephone other research centres to find out whether they had been replicated or could be explained. In

pharmacy the degree of specialisation meant that the number of outside groups would be extremely small, and *WF* suggested that there might be around 10 people in the world who might be considered experts on his own specialist compound. On the other hand, the cancer researchers associated themselves with national charities i.e. with their own source of funding, while the pharmacists looked to Germany and USA for related research groups in universities and industrial sites, and recognised that these countries had a large number of fields which were new and could offer them some kind of exchange.

Survey question 5) What are the main sources of information for your research?

Researchers in the sciences notoriously skim and scan their texts, often using them indexically (as we see below). The range of sources is therefore wider and more likely to be driven by indexes, both the basis of traditional indexes or on computer. Text books appear to be given much less priority, although they are obviously important for teaching (not a priority in the PSD). Research articles, indexes and electronic indexes were cited as primary information sources. Researchers were asked to select five journals of general interest and five that they considered essential to their own field. They found this rather difficult, presumably because of the sheer number of possible responses. Among the journals researchers mentioned, *Nature*, the *British Medical Journal* (BMJ), the *Lancet* and the *International Journal of Cancer* (IJC) were mentioned by over five researchers. *Science, Pharmaceutica Acta Helvetica* (PAH), the *British Journal of Pharmacology* (BJP), *Cancer Chemotherapy and Pharmacology* (CCP), *Cancer Research* (CR), *Journal of the Chemistry Perkin Transactions* (JCPT) and *Journal of the American Chemical Society* (JOACS) were all mentioned more than once.

Researchers also mentioned extensive use of the electronic Title and Abstract databases *MEDLINE, SCI, Index Medicus* and *ADONIS*. Some claimed that these were beginning to replace traditional 'journal loyalties' since a relevant title may be found in an index which covers hundreds of journals, all from the researcher's office. *PRL* suggested that regional and specialised journals would flourish since their coverage could be made more widely available through publication in indexes.

Survey question 6) In a given research journal, what criteria determine which articles are of interest?

There are central research articles and peripheral ones, and researchers clearly adopted different reading strategies once a decision of relevance had

been taken. Nystrand's dynamic reading model (1988) proposes that such decisions are probabilistic, based on factors that are given different weightings which change according to how far along the decision making process the reader has gone. Researchers were asked to demonstrate with a journal at hand which articles would attract their attention: *JG* proposed that he read around ten papers per hour from as many journals. Other researchers stated that they read from one morning a week to 'every spare moment', in the library or on the train, and when they occasionally had to check for specific information in the lab.

Key terms in Titles, as well as compounds in formulae, recognisable diagrams and data formats are the first entry points and the first clues. The respondents stated that specialist entities (a term I use later but first employed by *WF* when talking of specific compounds, cell lines, diseases etc.) were the main criteria, followed by or in combination with abstract properties or processes (stability, expression, total synthesis). Both entities and processes were inferable from titles, figures and reaction schemas, as mentioned in the introduction. Neither had to be exactly in the researchers' first list of major concepts. Another motivation for reading papers was curiosity, to catch up with related fields, or according to *PL* 'keep up to date general science I should know'. *DP* stated that a half-relevant term would 'fish out a subset' to provide a relevant connection. *WI* states certain preliminary questions that the researcher brings to the journal:

1. What things does it deal with?
2. Has anyone done this before?
3. Are there surprising results?
4. Do I believe it or not?

According to *WI* these would then lead to specific parts of the research article. In *MT*'s case, surprising results may be indicated by the number of animals used in the study and other methodological details. *PL* suggested that belief in the data was an important criterion: 'would the drug work with real patients?' *AG* stated that the main criterion for him was whether the paper offered a new model or alternative methodologies, not just providing positive or negative data. Several mentioned the *Journal of the Chemical Society*'s instructions for authors (1993: xii), which gives detailed rules on what is to be defined as 'new'. Among other rules: a compound is new if it has not been prepared before, if it has been prepared but was not adequately purified or was purified but not adequately characterised. Thus novelty must be judged in terms of claims against increasingly specific areas of other scientists' research.

The criteria of relevance are presumably different in electronic indexes where an initial stage of filtering precedes the processing of titles. *DP* gave sample figures of the kinds of titles he gets from the electronic index Medline. Of 300 titles from a 6 month period, he estimates that 150 will be already known, 100 useless and perhaps 3 or 4 on his specific area. The process of narrowing down in an automatic index (from the general key word *cancer* for example to *bacteriology*, or *cachexia*) appears to be more restrictive than reading entire titles in a journal where an entire proposition (sometimes in the form of an active clause) must be processed. In the journal, there is a chance that the title can be relevant (because of originality or peculiarity) without mentioning any specific keywords. This problem has been addressed by the SCI's Permuterm index, (SCI 1993) which accepts not only one word input but also entire phrases. Permuterm uses a hierarchical structure of key words and their phraseological or terminological synonyms (cancer, tumor growth, metastasis, oncology), followed by subject-specific collocations (such as *advanced, anorexia, associated, clinical*). Some semi-stop words (such as methods, analysis) are consulted only when key terms are identified. As in Phillips' (1985) study, high frequency words (full-stop words) are eliminated from the search, while other interesting middle-range terms are also eliminated (e.g. studies, consisting, shown). This classification of words implies a redundancy of high frequency items in indexing. However, the possibility of high frequency items being associated with rhetorical and phraseological patterns in the corpus does not appear to have been explored.

Survey question 7) What information do you derive from titles, abstracts, and other sections of the research article?

This revealed perhaps some of the most interesting discussion with the expert informants. Two reading patterns emerged: browsing and consulting. While *browsing* involves skimming the text for relevant details, *consulting* involves what I term the 'indexical' function: researchers use a number of different entry-points (graphics, keywords, bibliographic references) to approach the text. The text therefore becomes non-linear, and is structured accordingly to allow for this. Most generally, indexical reading takes place in the lab, when a straightforward fact is required from a text book or an index. The fact that some technical research articles are used in this way constitutes a major difference with research articles in the humanities, for example, and implies radical differences in the way the text is organised. Most chemistry texts for example establish temporary codes for relevant chemical compounds which allow the researcher to look directly at diagrams and then jump straight into the text. The information derived from different parts of

the article therefore depends on the expectations and expertise of the researchers and on the graphic properties of the text itself. It follows that not one part of a chemistry text can assume that the reader has read the previous sections, and much of my corpus is made up of repetitive, but linguistically interesting recapitulations. It appears that the more experienced researchers have more motivation to browse or read articles all the way through: *MT* claimed that he always checked the entire article, *PRL* claimed that he browsed 'more than the youngsters', while the (younger) pharmacists claimed that they read only partially.

Discussing how he dealt with titles and abstracts in journals, *DP* said that the decision to read on depended on whether the title was at the periphery or close to his field and how much he could derive from the Abstract. If a title or Abstract is on the periphery, DP looked up the rest of the paper only if there was not enough evidence in the Abstract. If there was sufficient evidence in the Abstract, he was content to take it at face value and to move on elsewhere. If papers were closer to his field, *DP* would 'glide through the article', focusing on the major findings if he couldn't explain them from the Abstract. Similarly, *PRL* claimed that familiarity with a field meant that the amount of attention and reading time could be reduced in the rest of the article: 'if you are clever enough you can infer the whole article from the abstract'. Thus partial reading is not indicative of irrelevance or lack of effort but simply the researcher's confidence in imposing a coherent reading of the text. The kind of information researchers expected in Abstracts and other sections closely resemble Swalesian moves. *PRL* claimed that an Abstract had four main elements in relation to the main article:

1. Inform the reader what it is about,
2. Tell the reader what you do in the paper,
3. Say whether you've succeeded in doing that, and
4. ('a bit of a luxury') Give future possibilities.

The role of the Introduction in the reading process appears to be ambiguous. Given the graphic nature of pharmaceutical research articles, their indexical use, and the relatively basic nature of the information in the Introduction, this section might appear to be redundant. Researchers spoke of the Introduction in terms of formally proposing and justifying current research. Others said that they expected to find the development of ideas presented elsewhere. *DL* stated that the Discussion section was the most important section for the reader, as it summarised the current research as well as suggesting or predicting an extension to the research model.

The pharmaceutical scientists (*SF, WF*) confirm our discussion above regarding the linguistic properties of Methods and Results sections. They

claimed that there was an overlap between them, as Methods sections start off as lab book transcriptions combining a template of measurements, while the Results 're-ordered' the measurements. This corresponds with an unexpected symmetry in my corpus: all of the 'Experimental' sections occurred in chemistry journals, and these often replaced Methods and Results sections in these journals (especially in the shorter 'communications' papers). Presumably the experimental data for the pharmacists can stand alone, while the shape of the data and medical applications can be treated separately in the Discussion section. In contrast, the microbiologists (*PL, MT*) saw Results and Discussion sections as distinct from Methods. Indeed, in the corpus all the joint Results/Discussion sections occur in microbiology and cancer journals. *PL* stated that this was because experimental data are seen as an 'extension to the research model' (as *AG* implied above) and thus in microbiology actual results should be interpreted and integrated in the context of medical applications.

This implied distinction between applied biochemistry and theoretical chemistry may be an oversimplification, but any distinction between these two essentially different positions means that not all of the rhetorical sections are equivalent, even if they have the same subtitle in different journals. As far as the corpus is concerned, this forces us to down-play some of the distinctions to be made between such sections as Methods / Results and Discussion sections. In practical terms, I was also obliged to exclude a small number of hybrid sections (most notably Results / Discussion sections) from the main Wordlist comparison, since the two sections were completely merged in some journals.

Survey question 8) At what levels do you write or otherwise contribute to the field?

Naturally, the most experienced researchers contributed in numerous ways (*MT* cites books, review articles such as the TPS article, book reviews, work in progress papers, *DP* cites seminars, industrial reports, international workshops), while everyone was involved with grant proposals, internal project reports and research articles (considered to be at the same level of prestige). This question was accompanied by a request to donate a published research paper for use in the corpus. For a discussion of the different types of research article obtained, see section III, 6.4, below.

Survey question 9) Details of writing up.

a) At what point of research does the writing of an article occur?

MT suggested that cancer research publication was essentially 'news oriented' - in the sense that as soon as a coherent story emerges from the data then it is worth publishing. *JG* (whose chemical processes actually use the metaphor 'stories' as a technical term) stated the same: writing up occurs 'when a block of information constitutes a story'. This was also the case not just for positive results but also for 'half-positive results', where there is a significant contradiction or difficulty to relate to the discourse community. As a chemist, *JG* writes data-oriented communications which, he claims, take a day to write but over a month to edit and redraft after discussions with colleagues. *WF* suggested that some writing up takes place before experimentation. This is presumably enabled by the serialisation of papers, and the template-like nature of experimental sections. Presumably researchers judge their own 'newsworthiness' in much the same way as they decide to read others' research papers, by centrality to a perceived problem, originality, and so on. Departmental factors must also play a part, and these may include peer-expectations, contractual obligation and inter-institutional competition for drug patents, which appear to be a particularly fierce area of competition in the pharmaceutical sciences.

b) Who is responsible for writing up and for editing?

SF and *WF* stated that if a research article is jointly written in a team, as are most of the papers in the corpus, different researchers take responsibility for different sections, with the central sections (note the use of the term 'central') such as the Experimental or Methods sections being built up by many individuals over time. This does not apply to the more experienced researchers, who either publish alone or, as *MT* and *AG* admitted, arrange for their research assistants to do the main writing up while they edit and correct.

c) How is the writing related to the research activity, and where is it stored?

Research articles are not only read in non-linear fashion, their production appears to be non-linear as well. Myers (1990) suggests that a paper is built and redrafted by several writers from the 'middle' out towards the Introduction and Discussion sections. Different members of the PSD conferred that they record reaction details of syntheses and other measurements over a period of months in the lab book with its various sections:

1. -Title (of extreme importance to avoid confusion of data)
2. -Date (to avoid repetition and to measure stages of progress)
3. -Reaction name
4. -Structural formulae (materials involved listed in shorthand codes)

5. -Reagents (catalysts and added materials for synthesis)
6. -Procedure
7. -Structural analysis of final product (in molecular percentages)
8. -Specific measurement details (yield, melting point, optical rotation, refractive index, elemental analysis...)
9. -Purity (checking contamination)
10. -Proof of structure (by blot analysis, NMR spectroscopy etc.)

This template provides the shape of the Methods, Results and Experimental sections. When transferred to the word processor, this list forms the backbone of the research article that can be fleshed out by adding explanations of unfamiliar procedures.

Survey question 10) What procedures exist to ensure the quality of research writing?

This question attempted to raise issues of editing as well as peer-review. All the researchers referred to the instructions for authors included in most journals. *The Journal of the Chemical Society (Perkin Transactions)* stipulates the format and the constitution of the research article, especially concentrating on the Experimental section and on the organisation of material (reaction schemes, the use of italics for position-defining prefixes, hyphens for chemical bonds etc.) as well as setting out rules for the authentication of novel compounds, this being the primary objective of the specialism. Contributions are generally judged on criteria of:

1. Originality of scientific content and
2. Appropriateness of the length and quality to content of new science. (*Perkin Transactions,* 1993: vii)

Echoing the kind of re-editing examined elicited by Myers (1990), the researchers confirmed that research articles have to undergo on average three or four re-writes before the final version is accepted. *MT* stated that editors generally correct structural aspects of papers, tone down claims and question the 'generalisability' of experimental data. Other researchers mentioned problems style. *MT, PRT AG* and *WF* all stated that the majority of editing deals with changes of emphasis and poor style, while *PRL* was also concerned that corrections of his own style appeared to be arbitrary and go 'unpunished' in other publications. Although 'grammar' and 'style' are mentioned by almost all the researchers as areas that consistently require correction, they were hard pressed to cite actual examples. *DP* was aware of standard procedures of politeness and for professional attack, including the damning: *it is rather surprising to find that x failed to find y* followed by a

proposed explanation, 'if you're feeling charitable'. *PRL* suggested that several clichéd phrases should be avoided, such as *typical results show that* and *preliminary experiments were conducted*. Several researchers claimed that their main problem in editing remained ata basic grammatical level, and there is some evidence that repeated structures are seen as poor style (*PRL* explored the possibility of eliminating the passive, for example, and replacing it with the imperative, as in cooking instructions!). Despite these reservations, it seems however that this phraseology resembles some of the most frequent and consistent expressions in the corpus. In addition, *SF* and others were surprised by my questions on repetition in articles. While they are aware of general stylistic constraints and general rhetorical functions, the researchers were often unaware of the role of reformulation and paraphrase in their texts. I asked *WF* and *SF* to talk through their papers in terms of the main message in each section, and they agreed that an important function of the various sections was not only to demonstrate methods and evaluate findings, but also to reword and re-explore concepts that had already been introduced elsewhere in the article.

III. Collocations and the Corpus.

In the first part of this book, I have demonstrated some of the complexities of the terminology and discourse of cancer research. In this section, I set out the theoretical and technical notions of phraseology and collocation on the basis of Firth's theory of meaning. This prepares the way for an analysis of collocations in research articles in section IV. As collocational analysis requires large amounts of authentic textual data, the final sections of this section set out the design features of a representative corpus of cancer research articles: the Pharmaceutical Science Corpus (PSC).

1. Choice in the Grammar of Texts.

It is relatively straightforward to describe the linguistic features of scientific texts. The computer enables us to identify large numbers of regular expressions, and a well-designed corpus analysis should be able to automatically recognise given linguistic features as the typical style of a specific genre or type of text. The main issue however is not our ability to spot long-term patterns, but the extent to which we are able to identify relationships between these expressions and their relative value when used in a real text and by a real scientist. And although Chomskyan and generative theories of language have proven to be valuable models of potential expression, mainstream linguistics does not provide us with the conceptual apparatus necessary for a description of style within a particular discourse. I propose here that the analysis of collocation presents an ideal opportunity for such discourse analysis. However, it is important to be able to situate isolated examples of collocation within a broader system and to explain their significance within the discourse of science. What is needed therefore is a linguistic account of choice of expression, and it is for this reason that many descriptive studies refer to Firth's ideas on language. As Firth was also the first linguist to place the term 'collocation' within a theory of meaning, an overview of his theories of language, and their development in Halliday and Sinclair's work are central to a theory of collocation in general.

Apart from the concept of collocation, as noted in the overview (section I.3), the main contribution of Firth has been to argue that there are many levels of meaning:

> ....the complete meaning of a word is always contextual, and no study of meaning apart from a complete context can be taken seriously. (Firth 1935:37).

Here 'context' refers to textual context (co-text) in the first instance, but also to semantic knowledge and Malinowski's 'context of situation'. The point is argued in similar terms by Wittgenstein, who not only conflates meaning with use, but also links our understanding of an instance with our knowledge of the whole system:

> The meaning of a word is its use in the language...To understand a sentence means to understand a language. To understand a language means to understand a technique. (Wittgenstein, 1957, ¶199)

Firth's 'polysystemic' principle is therefore based on the structuralist idea that 'if a new term is added to the system this changes the meaning of all the others' (Halliday's reformulation: 1961:247). Firth suggested, for example, that the meaning of the nominative case in a two case language would be qualitatively different to its meaning in a four-case system (1957:190,227). Although the linguistic form of the nominative is the same in both systems, its underlying meaning is altered. The same is presumably also true between varieties of the same language. Because the distribution of grammatical resources varies from one variety to the next, the underlying meaning of a given grammatical feature changes according to the system it is currently engaged in. By primarily defining linguistic terms as functions, Firth thus appeared to undermine the usual practice of linguistics which was to see form as the primary basis of definition.

In the case of science writing, we have seen that the underlying meaning of the passive, of forms of nominalisation and the use of modal verbs is extended and modified by their use in the specialist language, and that these uses (and therefore meanings) are often at one remove from their equivalents in other varieties of English. Both Halliday and others (for example, Banks 1997), explain many of these functions in terms of abstraction, hedging and grammatical metaphor (discussed below). These function-labels cut across the boundaries of form. And as we have seen in section II.3, Firth's polysystemic principle underpins Swales' concept of 'discourse coherence'. This perspective leads us to distrust the notion of sublanguage and other characterisations of texts which rely on single grammatical features, or

ascribe to a single feature a functional role which remains constant across a series of genres or registers (as in Biber's 1986 multifactorial technique).

A problem with all functional grammars lies in the extent to which it is possible to map discourse functions to language forms. Halliday approaches this problem from Firth's perspective of 'modes of meaning'. Halliday (1985) suggests that choices of expression are not isolated and simple but involve simultaneous decisions involving three basic **metafunctions**. The notion of metafunction emerged from Halliday's work on intonation and variable emphasis of mood and theme in spoken English. Halliday noted that intonation in the spoken language is used to great effect in English, and allows the same sentence to be modified according to its propositional meaning, thematic focus and rhetorical force (Halliday uses the terms *ideational, textual, interpersonal*). The written language clearly requires these functions as well, but must express its 'intonation' with different resources: for example, by a more complex form of syntax (hypotaxis, embedding) or by signalling emphasis graphically (by capitals, exclamation marks, italics, paragraphing, punctuation etc.).

Halliday proposes that language varieties realise the three metafunctions in different ways. This can be demonstrated using a single example from the discourse of science:

1. This protein is thought to be a major factor in breast cancer.

The ideational function corresponds to the traditional view of transitivity as an expression of participant, process and circumstance (Halliday and Hasan 1989:68). In the example sentence, the subject of the verb *this protein* is a 'participant' but is not felt to be the agent or initiator of some action - the usual function of the grammatical subject. Instead, *this protein* represents a 'token' which is attributed a value expressed in rest of the clause (*a major factor in breast cancer*). Ideation is therefore a purely semantic relation within the clause.

The textual function takes a different perspective, and involves the way the message is presented in the surrounding discourse. For example, science tends to organise its messages by constant reformulation. In the sentence above, *this* is used to encapsulate and refocus a previous discourse topic (a *protein* – a backwards reference to a complex chemical compound). This is a lexical reformulation and tends to involve a more general word or a new formulation with some degree of evaluation (*This unusual orientation indicates that ..., This surprising result prompted us to ...*). Thus while the ideational function emphasises grammatical roles within the message, the textual function relates the message to the running text. The textual function

is typically seen in the use of the passive. From a thematic point of view, the passive effectively 'saves' new information in the message until the end of the sentence. Although this is seen as a prototypical feature of science writing, the same process occurs in other genres, especially news reporting (McCarthy and Carter 1994).

Finally, the interpersonal metafunction involves the clause as a rhetorical proposal which can be subjectively asserted or qualified. In science writing, the interpersonal function is realised by various impersonal devices which effectively obscure the direct involvement of the scientists or express some degree of 'polite' hesitation in order not to overstate the claims of the author, as pointed out by Myers (1989). Modality in science involves inanimate subjects (*results suggest that*), the hedging of data using modals (*it may be the case that*), the use of mental or verbal process nouns (projecting nouns such as *belief, suggestion*) and, as might be expected, the generalised use of the passive (*cell growth was analysed*). In the above example, the sentence can be seen to have the same propositional meaning as *This protein is a major factor in breast cancer,* but incorporates a further degree of modality in the form of a mental process verb (*thought*). This is further modalised by a passive (*is thought to be*) in contrast to a more direct alternative '*we believe this protein to be a major factor...*'.

Thus from Halliday's point of view, a specific grammatical form can be treated to different kinds of interpretation within the same overall framework. The passive emerges as a simultaneous collaboration of three different choices: a way of placing the agent or medium (an ideational function) in the 'new' position of the clause (a textual function) at the same time as avoiding the expression of personal involvement (an interpersonal function). Although the metafunctions are often discussed in terms of clauses, they are not tied to grammar alone and have provided a framework for lexical studies of idiom (Fernando 1996) and the analysis of scientific texts (Wikberg 1990, Mauranen 1993).

The concept of value-related choice is at the heart of Halliday's systemic grammar. As Halliday puts it:

> The system of available options is the 'grammar' of the language, and the speaker, or writer, selects within this system: not *in vacuo* but within the context of speech situations. Speech acts thus involve the creative and repetitive exercise of options in social and personal situations and settings. (Halliday 1976:142)

The term 'systemic' therefore indicates choice within a system. The concept of choice does not imply free expression with infinite possibilities, but instead indicates a continuous spectrum from a typical to a more marked

expression. Halliday (1991,1992) therefore proposes that choices in functional grammar operate on a probabilistic basis. He suggests that closed systems in language oscillate between equiprobable systems (past vs. non-past tense, singular vs. plural) and systems which are skewed (affirmative vs. negative polarity in the clause, passive or active voice). Equal probabilities in the system are likely to indicate a largely redundant choice, whereas skewed probabilities assign a high level of emphasis on the infrequent or marked choice.

In a pilot study designed to demonstrate this hypothesis, Halliday and James (1993) examined 25 high frequency verbs in an early version of the Cobuild corpus (20 million words). They found that clause polarity is distributed at a ratio of roughly 90% / 10% while the primary tenses are distributed roughly equally (50% / 50%). These probabilities are then assumed to vary according to variety of language. In science texts, Barber (1962) observed that of 1770 verbs observed in astronomy, biochemistry and electronics, 89% are in the simple present and 11% in other tenses. The past tense is therefore marked in scientific articles, but in the language as a whole (where it is equiprobable) it represents an unmarked choice. Barber also found that the active / passive voice was slightly less skewed than normal at 65% / 35% and thus represents a less marked choice. In Halliday's model of register therefore, words and grammatical constructions have an inherent probability attached to a specific discourse or register. As Halliday says: 'frequency in the corpus is the instantiation ... of probability in the grammar.' (1992:66). Such a system of probability in the grammar has important implications for the interpretation of statistical results in the corpus, as Sinclair notes (1993c:167).

Nevertheless, these probabilities are not fixed properties of specific varieties: we still have to account for Swales' 'discourse coherence', and the possible recasting of stable grammatical features into new roles. When Halliday refers to the **register** of science, his definition avoids explicit reference to grammatical form as a constant feature, and he instead prioritises the favoured status of certain forms from the point of view of the system as a whole:

> [Science writing] is English with special probabilities attached; a form of English in which certain words, and more significantly, certain grammatical constructions, stand out as more highly favoured, while others correspondingly recede and become less highly favoured than in other varieties of English. (Halliday 1993:4)

This view is not far removed from Enkvist, who provided a definition of style that is tailor-made for corpus linguists, being statistical in nature as well as incorporating the idea of register change:

> The style of a text is a function of the aggregate of the ratios between the frequencies of its phonological, grammatical and lexical items, and the frequencies of the corresponding items in a contextually related norm... past contextual frequencies change into present contextual probabilities, against whose aggregate the text is matched. (Enkvist 1964:28)

It is perhaps useful then to conceive of a register as a variety of language in which all the resources of language are still available but are marked for use as 'central' or 'peripheral'. This notion is perhaps more flexible than that of the 'sublanguage' (see Chapter One), which does not distinguish between the core or peripheral features but situates the sublanguage *as a whole* in relation to other sublanguages.

In support of the Hallidayan perspective, many studies have shown that the grammatical features of registers are historically contingent and open to free variation (Atkinson 1992). There is no such thing therefore as a prototypical language of science or a fixed set of grammatical features, but instead a series of Wittgensteinian 'family resemblances', features which come into focus or fade away as the register moves in time. Registers are thus inclusive of the whole language system, and any linguistic resource, no matter how marginal, may undergo a revival within a specific discourse. Halliday and Martin expand on this idea in their discussion of the historical development of science writing. They claim that as a society changes its system of self-expression, existing linguistic resources take on new roles (1993:9). Halliday points to the fact that whenever there has been major social, political, or technological upheaval, there have been shifts in the use of language. Thus nominal expressions were introduced in medieval Latin to deal with the philosophical and administrative tasks of the new written language. The renaissance and the industrial revolution were in turn landmarks for linguistic change in the major languages of science, particularly French (Lodge 1996).

Halliday and Martin show that the same processes are still evolving in technical and scientific English, in particular in the role of nominals. Martin (1991) points out that an important function of compound nominals (for example, *cancer patient, cell growth*) is to state a specific argumentation as 'given' or 'understood'. For example, the grammatical relations between 'cancer' and 'patient', and between 'cell' and 'growth' are not expressed. It is only in an extended grammatical paraphrase or reformulation that the relations become more salient (*patients with / suffering from cancer*, *growth*

from the cell / the cell as it is grown). Halliday has shown that scientific texts systematically construct compound nominals by building the nominal up piece by piece until several explicit grammatical relations are finally hidden. The following example demonstrates how compound nominals are typically formed within in a single text (Halliday 1992:70-71):

How glass cracks ...	The stress needed to crack glass ...	As a crack grows ...
The crack has advanced	... will make slow cracks grow	The rate at which cracks grow ...
The rate of crack growth ...	We can decrease the crack growth rate ...	Glass fracture growth rate ...

Although nominalisation of this type allows information to be reformulated with greater flexibility within the clause, the underlying propositions in the compound become increasingly difficult to interpret or de-construct. Once formed, compounds may tend to become idiomatic and to some extent beyond interpretation on the basis of individual elements. While the lay reader may be able to guess the meaning of a nominal such as *glass fracture growth rate*, it would be impossible to meaningfully explain the term and the relations between each element without reference to the original text. In other words, as Halliday states, the meaning of the compound is 'instantial', couched in the text itself. This corresponds to the creation of new terms and collocations – and we can see in this example a clear case of collocation and terminology being created as a natural product of a text. Halliday terms this 'logogenesis' (1992:70) and it seems that few works on terminology, with the exception of Pavel (1993a), have emphasised the primarily textual creation of terminology.

As I have noted above, the terminology of science is often seen as rationally planned by groups of experts rather than emerging from a single text. Halliday and Pavel have shown however that texts are instrumental in terminological innovation. In addition, the compound nominal pattern has been recognised for some time as an important feature of scientific English. Such is the pervasive nature of English phraseology that languages which do not normally favour the juxtaposition of nominal elements without a preposition or other relational marker (including French and other Latinate languages) are beginning to adopt this pattern from English, most usually in their technical and scientific terminology (Bauer 1979). Similarly, Stubbs (1996) has pointed out a parallel evolution in English verbs, namely the increasingly ergative use of verbs such as *'the bank closed*' and *'the factory shut down*'. Stubbs claims that the ergative function is symptomatic of a

general discourse in English which obscures the nature of agency. His conclusion is that ideology is implicit in linguistic choice, to the extent that the selection of even one feature from a set of alternatives is indicative of some rhetorical intent. We should note that it is only since the advent of large-scale corpus analysis that such grammatical tendencies have been open to systematic examination.

Halliday sees 'ergative verbs' and nominalisation as instances of a more general process in scientific discourse: **grammatical metaphor**. Traditionally 'metaphor' is associated with a lexical transfer or allusion. Grammatical metaphor consists instead of the transfer of information from one grammatical role to another. Halliday uses the following example: '*the fifth day saw them arrive at the summit*' as opposed to '*they arrived at the summit on the fifth day*'. In the first example, *the fifth day* becomes the grammatical subject but functions semantically as a metaphorical observer. Several linguists have observed the effects of grammatical metaphor in science writing. Banks (1994b) examines the use of research-oriented verbs with inanimate subjects, as in *The current meter at mid-depth [...] provided data... [This] photographic technique will produce underestimates of abundance*. Banks compares these marked expressions with the general language where inanimate subjects are the privileged subjects of events (*Water flows* rather than a marked event: *Geoff flows...*). He concludes that grammatical metaphor is a major linguistic resource for obscuring agency and authorial responsibility in scientific writing.

From these observations, we can conclude that the traditional preoccupation with the passive, the quintessential feature of impersonal style in scientific writing, has to some extent obscured other fundamental features of language which are equally central to scientific thinking. Ergative expression of verbs and nominal reformulation are both realisations of a common function in science, the 'impersonal style' identified by Biber et al. (1998). But they also have a fundamental role in the textual expression of ideas, a point that is difficult to identify from a statistical word-count. Although the passive is an easily identifiable feature of written science, it is clearly only part of a wider system and we need to bear this in mind in our analysis of the corpus.

I have mentioned at several points that reformulation is a key process in the development of scientific ideas. While collocations have not usually been analysed in terms of their role in the text, a number of studies have argued that lexical items and the lexical system as a whole may have an important role to play in our understanding of text and text structure. Halliday and Hasan's model of **cohesion**, defines text as a series of explicit relations that

distinguish it from a random string of sentences (1976:6-7). The essence of the cohesion model is that grammatical reformulations (such as elision, substitution and pronomial reference) as well as lexical items are seen to have a role to play outside the traditional syntactic boundaries of the sentence: either by signalling links outside the text (exophora) or backwards and forwards beyond the level of the sentence (endophora). Lexical cohesion involves reiteration and reformulation of items throughout the text, the use of synonyms or superordinate words and a broad form of collocation (1976:278). (Halliday and Hasan's collocations are items which 'share the same lexical environment' such as *doctor* and *clinic*, i.e. a paradigmatic relationship as well as a syntagmatic one 1976:286). Thus grammatical reformulation and lexical items not only have syntactic relations within the sentence, they also represent choices that are cohesive in nature and serve to signal relations within the wider development of a text.

On the basis of Winter's (1977) work on lexical signalling, Hoey (1983) analysed the distribution of lexical cohesion in text. He found that lexical cohesion was of wider importance in the text and of greater complexity than the other more traditional categories of cohesion, such as conjunction. He argued that the role of lexis was crucial in textual organisation, so much so that almost every lexical choice in the text could be seen as an 'encapsulation' or 'prospection' of ideas in the surrounding co-text (terms proposed by Sinclair 1981, 1993b). Words are therefore not simply selected as collocations or syntactic constituents in the clause, they are constrained and interpreted within the running text. This observation clearly contradicts the traditional view of writing, which sees 'discourse markers' as the main elements in the organisation of the text (Hoey 1983:176). Hoey proposed instead a 'non-linear' view of discourse. While signalling of all types clearly aids the explicit formation of a coherent text, Hoey argued against the traditional view that texts are set out in an implicit dialogue between writer and interpreter, and instead predicts that discourse is built up of incomplete and unfinished texts (1983:177):

> We are all contributing to one interwoven discourse, of which our own contributions are but incomplete fragments. (1991a:159)

This militates against the view of a text as a unit where every semantic signifier and signal plays an equal and necessary role. Hoey's conclusion is that texts may make use of fixed expressions in order to allow the reader to predict content and argumentation (1991a:154). He points to cloze testing where informants successfully fill in lexical gaps and reconstruct coherent text (he calls this the *Jabberwocky* principle, since the only clues lie effectively in identifying the typical members of meaningful grammatical

frameworks). This may also explain the observations I set out in the survey, which suggest that researchers read 'indexically'; that is, they are able to successfully predict and by-pass much of the linear detail of the research articles they have to process. As an extended reformulation, the research article need not be read from beginning to end for all purposes. Lundquist (1989, 1992) appears to provide evidence for this by showing that non-experts who read scientific texts tend to rely heavily on lexical networks to establish long-range links, while experts do not need explicit signalling and are thus able to skip and skim through the text and establish a meaningful but partial reading of the text (1989:141).

However, Myers (1991) has argued cohesive systems are in fact specific to different registers, and take on different functions in the research article genre. In his analysis of cohesion in science writing, Myers (1991:13) points out that a reliance on lexical networks is not enough for non-expert readers. Myers underlines the difficulty involved in deciding how cohesive lexical repetitions really are, especially in terms of synonyms (*DNA* vs. *genome*) and superordinates (*molecule* vs. *product of transcription*). He argues (1991:5) that background knowledge of the scientific paradigm is essential for any networks to be built up, and this accounts for the differing forms of cohesive devices used in scientific and popularised texts. As with Hoey, he suggests that phraseology may be the key to understanding cohesive relations:

> Some cohesive devices depend on the reader recognising collocations, and using them to unpack dominance relations in noun phrases. (Myers:1991:14)

This observation brings us back to Halliday's work on grammatical metaphor and the reasons why scientific texts are written in such a specific style. It emerges from our discussion above that scientific research articles are not only a series of arguments linked by progressive reformulation, they are also non-linear indexes, allowing scientists to approach the text from several entry-points and to use fixed expressions and lexical cues to orient their way around the text. From a traditional perspective, these very specific properties of science writing might be considered irrelevant to the stylistics and syntax of the text, but in a Hallidayan grammar they are considered to be determining features in the lexico-grammar and phraseology of the genre.

As with many aspects of Halliday's writings, our discussion has led us to an examination of specific examples and then on to a proliferation of more theories. As de Beaugrande has pointed out, Halliday makes no attempt to reduce grammar to a uniform minimal structure, but instead '[his] grammar enables an analysis in which richness and multiplicity steadily increase'(1991: 258). I have set out here some of the theory that has been inspired by the work of Halliday's work on choice in text. This leads us now

to discuss Halliday's notion of lexico-grammar and then move on to Sinclair's related notion of the idiom principle.

2. The Lexico-grammar

In the Introduction, I set out some of the theoretical issues surrounding the notion of collocation, and suggested that collocations can be analysed in terms of three increasingly complex standpoints: *statistical / textual*, *semantic / syntactic* and *discoursal / rhetorical*. I argued that these three perspectives are compatible and bring considerable value to the notion of collocation. The statistical / textual approach insists on collocation as a product of on-going discourse and seeks data which is unconstrained by theory and categories which may be 'self-selecting'. The semantic / syntactic approach on the other hand demonstrates the need to restrict the analysis of collocation to meaningful expressions and the need to establish the internal cohesive properties of each phrase. Finally, the 'discoursal / rhetorical' perspective underlines the textual function of collocation as well as the idea that collocations operate in a system of alternative choices of expression. It is not surprising that the three approaches lend themselves naturally to a three-stage methodology (data analysis, data selection, interpretation), and I attempt to set this out my corpus methodology, below.

While I demonstrated that there are several ways of identifying collocation, they still remain abstractions and far removed from actual processes of data collection and analysis. Here I argue for a particular focus, the analysis of grammatical items in the corpus. This is based on the belief that an untagged corpus needs to be analysed in a systematic way. In addition, some research, especially Sinclair (1991) indicates that grammatical items can provide a useful way of initially approaching a large mass of data. Grammatical items appear to be excellent indicators of general phraseology, yet they have not received as much attention in general lexicology or corpus linguistics as their lexical counterparts.

The irony about grammatical items is that although they happen to be extremely frequent - and therefore from a Hallidayan perspective, extremely important – they also happen to be *too* frequent. So much so that they are usually systematically eliminated from statistical counts, especially in large scale textual analyses, where the researchers are forced to concentrate on lexical collocation (Phillips 1985, Smadja 1993). Workers in information retrieval and automatic abstracting term them 'stop words' and happily describe how they are able to automatically extract them from an index or

data base (Luhn 1968, Yang 1986, Källgren 1988a and 1998b, Wilbur and Sirotkin 1992). Previous studies have claimed that high frequency items are stable in use and meaning across different types of language, and the reverse assumption is that if a word is stable it is a 'grammatical item' or a 'function word'. Sager et al. (1980:238), for example associate a descending type / token ratio (a measure of the density of different word forms) with increasing levels of specialism in technical texts, that is: the most frequent words in the language account for proportionally less of the total vocabulary of LSP texts. They assume from this that high frequency words are of little use in the analysis of specialist texts. Phillips also characterises grammatical items as noise, distinguishing them from 'carriers of local meaning in text' (1985:66). There are obvious justifications for this in an automatic analysis of semantic structure in text. The assumption of redundancy has also been applied to high-frequency items, even in collocational studies such as the *BBI* dictionary (Benson et al. 1986) which eliminates common words (such as *big, cause* and *make*). And the influential lexicologists Thoiron and Béjoint have stated that high frequency words can collocate with 'almost any words in the language' (1992:7).

Yet if we are to adopt a systemic approach to discourse, it is important to see grammatical items as fully part of the lexical system as a whole. While Halliday proposes a theory of grammar and Sinclair works on lexis, both view lexis as the bedrock of grammar and both see grammar and lexis in terms of a continuum rather than a categorical divide. Halliday in fact terms the complete grammatical system a 'lexico-grammar', where grammar is a heavily constrained and abstract form of vocabulary rather than a separate linguistic level:

> Grammar and vocabulary are not two different things; they are the same thing seen by different observers. There is only one phenomenon here, not two. But it is spread along a continuum. At one end are small, closed, often binary systems, of very general application, intersecting with each other but each having, in principle, their own distinct realization [...] At the other end are much more specific, loose, more shifting sets of features, realized not discretely but in bundles called 'Words', like *bench* realizing 'for sitting on', 'backless', 'for more than one', 'hard surface'; the system networks formed by these features are local and transitory rather than being global and persistent (Halliday 1992:63)

Sinclair's theory of lexis is embodied in the **idiom principle**. It is a provocative theory of collocation, in that it eschews many of the assumptions of mainstream corpus linguistics. Sinclair does not view tagging (marking up of the corpus) as essential, and analyses word forms without reference to the

lemma or base word. Thus *goes* and *went* are analysed separately from the base form *go*, as though they are separate lexical items. As we have seen in the previous chapter, Sinclair holds collocation to be a purely statistical and syntagmatic feature of language: collocations do not have to be fully grammatical, and are not necessarily limited to the boundaries of the phrase or the clause. And as with Nattinger and DeCarrico's approach, this feature alone makes Sinclair's idea of collocation a very different notion to the mainstream view in lexicology and phraseology studies.

The starting point of the idiom principle is that the collocational behaviour of a word is not an issue of individual item selection, but depends on the unstable and shifting nature of the word as a whole unit and the indeterminate nature of its grammatical class, at least in a historical perspective. Sinclair points to word blends as clear instances of items that have lost their status as separate words in English (*because, of course, maybe, another*, *altogether, alright* etc.). Many of these expressions represent the kind of grammaticalisation observed in the development of pidgin languages: the gradual formation of grammatical words from bound lexical phrases (Traugott and Heine 1991). For example, Tok Pisin uses the lexical *bye and bye* and *finis* from English as grammatical particles of aspect. Words are therefore not fixed in position but may be used along a continuum from pure vocabulary items to features of grammar. This degree of continuum from one category to another is also evident in in lexical paradigms. Hence suppletion is seen in forms such as *went* (originally derived from the verb *to wend*), which historically drifted into the paradigm for the verb *to go*. The conjugation paradigm of a verb may be a cognitive reality, but its constituents are historically contingent and unrelated.

This kind of long-term change suggests that the upper level boundary between the lexical item and the phrase is in constant flux. But there is also evidence for what might be seen as the development of larger-than-word lexical items in the contemporary language. Nattinger and DeCarrico (1992:24) and Willis (1993:88) refer to holophrastic phrases: prefabricated chunks of language which lead a clichéd or marginal existence, including *wannabe, allgone*, *watsup?* Similarly, high frequency content words (such as the delexical verbs *get, make, set*, *take*) also depend on complements or particles to be fully lexical semantic units (*get even, get on, make for, make way, set up, set off, take place, take part* etc.). Sinclair has suggested that the many combinations in which these words enter must form a large part of the total lexicon (rather than a simple count of single lexical items), and that many texts may be characterised as being largely 'delexicalized' (1987c:323). This modifies somewhat the traditional view of lexical density (Ure 1971), which relies on a count of lexical forms and does not normally

take account of grammatical items as part of longer or meaningful expressions.

On the basis of such evidence, it is possible to dismantle the traditional view of the strictly delimited word-class. Sinclair and his co-workers on the Cobuild dictionary have consistently emphasised the unique nature of single grammatical items, and their main argument has been that high frequency items tend to have unique lexical properties in comparison with the rest of their traditional word class. For example, the very frequent preposition *of* does not share the properties of other canonical prepositions in adjunct phrases or as the indirect complements of verbs (Sinclair 1991). Some high frequency lexical words are also seen to be 'grammaticalised', to such an extent that no two lexical words could be seen to have exactly the same collocational properties. At the heart of this view is the notion of the 'pattern' (Hunston and Francis: 1998): the idea that grammatical items and lexical items are chosen in tandem with a specific formula in mind rather than selected individually or 'compositionally'.

On the basis of a large scale corpus study of nominals in English, Willis (1993) has shown how classes of word merge into one another and how some subsets of the noun have very different properties to the traditional class as a whole. For example, only a subset of all nouns modify the semantics of delexicalised verbs (*give a smile, take a chance*) or are involved in projecting clause structures after *that* (*the belief that, the argument that*). This subset differs from those nouns which can take infinitive verb forms (*a decision to, the claim to*) or complex nominals with *of* (*behaviour of, arrival of*). Thus nouns do not all share the same collocational properties, and these 'families' are more specific and consistent than the notion of 'abstract noun', which is sometimes assumed to be a catch-all for nouns which become involved in complex phraseology. (I have also suggested that the distribution of these nominals between the different categories of noun is different in other languages - Gledhill 1999).

Willis also notes the rhetorical role of nouns in structures such as sentence stems: *the* (*main / important / other*) *thing is that... the (question / problem / difficulty) is that...* He argues that the main feature of these expressions is not the family of noun involved, but the fact that each entails (or collocates with) a further expression in the projected dependent clause, in this case a signal of some solution to a problem (1993:88). Rhetorical functions collocate therefore with specific nominal phrases. In a similar study, Francis (1993) has discussed the pre-emptive properties of what she calls semi-idiomatic phrases as in *put a brave face on it*, 'semi-pre-packaged' idioms with clear communicative goals (*not the foggiest / faintest idea*) or prefacing items (such as *is a case of*) where a current discourse topic is compared to one

familiar to the reader (1993:143-6). Altenberg (1991) has similarly argued that many collocations extend beyond the traditional bounds of the phrase, and are therefore not analysed in mainstream lexicology. He points to the cognitive sub-system of 'amplifier collocations' such as *absolutely* which occurs with superlative adjectives, and *perfectly* which collocates with positive and negative adjectives. The correspondence between grammatical form and semantic or discourse functions hardly seems to fit into a traditional paradigm of phrase structure syntax or feature-based semantics.

While the nature of the word-class and the word-boundary has been reassessed on the basis of corpus work, so has the relationship between grammatical collocation and more fundamental syntactic structures. **Grammatical collocation** traditionally involves the collocation of grammatical items with a limited set of lexical items (Howarth 1998:184). In her work on the Cobuild corpus, Francis demonstrates how a high-frequency pronoun (*it*), a conjunction (*that*), an adjective (*possible*) and a noun (*reason*) each have their own lexico-grammar, and interact with increasingly delimited forms of syntax (1993:140). Francis finds that *it* is likely to occur as a grammatical extraposed form in adjectival complement clauses: *they often find it difficult to explain why* or *they often find explaining why difficult*. Whereas a descriptive grammar might present this as a general pattern, Francis points out that the structure is limited to just two main verbs *find* and *make* (98% of all occurrences of extraposed *it*). The structure in turn collocates with a very restricted set of adjectives (related to the concepts of ease and probability) and to two specific expressions *make it clear / likely that*. Francis also finds consistent patterns (1993:46) for the adjective *possible*, which mostly occurs with superlatives in the frame *as X as possible* or after *whether / if...* Similarly, *that* in NP complement clauses (as in *the idea that, the advantage that, the chance that*) has a limited series of structures that can be classified semantically, such as illocutionary processes (*allegation that, contention that*) and thought processes followed by results (*analysis that, realisation that*) (1993:149). When used in NP complement clauses, the noun *reason* has two patterns: introducing an event (*the reason he fell... the reason I did this...*) and an expression of contrast with the collocation *'for the simple reason that...*' which introduces an explanation rather than an event (*for the simple reason that he was drunk, for simple reason that it was a good idea*) Francis (1993:147) concludes that grammatical items and syntactic structures (such as extraposition, complement structures and so on) operate selectively with a limited set of lexical items. Hence very frequent grammatical structures map onto consistent patterns of meaning. As with Hoey's view of lexical cohesion, Francis claims that collocations are chosen as a strategy of communication

rather than simply to express complex ideas in a succinct form. As Sinclair puts it 'grammar is part of the management of the text rather than the focus of the meaning-creation'. (1991:8).

The analysis of grammatical collocation has demonstrated that the boundary between grammatical and lexical items is a relative one. Sinclair and other corpus linguists have long argued that linguistic behaviour is not openly accessible to introspection and can only be properly examined on the basis of authentic text analysis. Native speakers are typically unaware of the collocational structures that are systematically found in computer-based corpora, and are certainly not able to guess the relative probability of one structure compared with the next. For example, Kennedy (1984) has reported that 63% of the use of *at* is limited to 150 collocations, with *at least* being the most frequent. Similarly, Krishnamurthy (1987:70) reports that many common items have very restricted collocations, such as the 70% co-occurrence of *refer* with *to*, while 100% of the uses of *encrusted* are as an adjective rather than a verb, and *backsliding* as a noun rather than a verb. Carter (1998: 197) has noted that these very consistent collocational properties and probabilities are significant evidence of lexico-grammatical competence, and lead to a more probabilistic view of a native speaker's mental lexicon.

However, not all linguists are happy with the corpus-based analysis of grammatical items. Moon (1987) has suggested that an emphasis on context, especially with high frequency words, has led to an over-abundance of meaning distinctions where, in lexicography at least, the analysts runs the risk of 'losing the semantic integrity of the word.' (1987: 102). She argues that the collocational analysis of grammatical items can not reveal much if the item happens to collocate with others at a distance, especially grammatical words which express discourse or clause functions (*and, but, however*) or collocations which appear to require quite a large cotext such as (*so ... as*) as Kaye (1990:151) notes.

While this is an important consideration, Moon's point is aimed at delimiting examples and establishing essential meanings for entries in a dictionary. If we are considering the lexico-grammar of a particular style or register, the corpus evidence, as we have summarised it above, appears to strongly favour the discussion of grammatical items and grammatical categories in relation to collocation. While discourse analysts may be tempted to conduct corpus analysis on the basis of lexical items, the notion of the lexico-grammar suggests that phraseology is of equal importance in the meaning-creation of the text. And as we have seen, an analysis of grammatical items can be used to 'trawl' for the fundamental phraseology of the text. Grammatical items are the starting point, but grammatical

collocation is not just simply about the grammatical items themselves. The theory of lexico-grammar implies that grammatical items are simply consistent elements in longer-range fundamental phraseology.

We have seen so far that a statistical analysis of collocation may be a sufficient basis for establishing the basic collocational properties of words. We have seen that grammatical collocation is an important feature of the general language, at least in English, and that certain studies have posited a fundamental role for collocation as a bridge between the notion of the word and the text. However in practice, as I have noted in the Introduction, the statistical notion of collocation needs to be restricted (in terms of the internal cohesion of the expression) and also requires a more contextual interpretation (in terms of its place in the general discourse). These issues are well known in the field of corpus linguistics and lead us to a wider discussion of approaches to corpus analysis and the identification of collocations in specific text archives.

3. Corpus Linguistics

Corpus linguistics involves the collating of linguistic features from a computer-held archive of texts, where the corpus is representative of some part of the language. The use of computers for data collection has not only entailed a massive increase in corpus size (from thousands to millions of words), but also a transformation in theories of linguistic description. Burnard (1992:2) states that this approach is so different from other types of linguistics that it necessarily entails the 'development of new, pragmatically derived linguistic models'. Leech (1992) similarly emphasises that many corpus analysts share a set of core assumptions which are not widespread in mainstream theoretical linguistics: an interest in the empirical, quantitative description of language in use. According to Leech, the main advantage of the computer-held corpus is that there is a sense of exhaustive or 'complete' use of data, as opposed to highly selective use of data in other linguistic fields (1992:112). A second advantage is the availability of 'test corpora' to quantitatively test findings worked out on other archives of texts. A corpus-based model of linguistic behaviour is therefore falsifiable because it can be tested against fresh data. At the same time, the text corpus can be distinguished from a text archive or reference-tool such as the *Trésor de la langue française.* The corpus allows for open-ended linguistic analysis (the archive limits the format of searches) and permits linguistic intervention (especially tagging) of the texts in the corpus. Corpus linguistics has built up a reputation in such diverse areas as speech recognition modelling (Church

and Mercer 1993), word association tests (Church and Hanks 1990), natural language processing (especially the application of syntactic notation: Leech and Fligelstone 1992), general lexicography (Clear 1987, Sinclair 1987), semantic labelling for dictionaries and language research (Vossen et al. 1986), machine translation (Schubert 1986), the development of terminological knowledge banks (Ahmad et al. 1991) and the development of language teaching materials and syllabuses (Willis 1990, Johns and King 1993).

Generally speaking, there are three different schools in English-speaking corpus linguistics. Firstly, there has been much corpus-based work in computational linguistics and terminology, with a long tradition of statistical modelling (Butler 1985a, Oakes 1996). Secondly, descriptive linguistics has concentrated on the tagging and parsing of corpora, usually within a generative framework (the Lancaster school: McEnery and Wilson 1996). Similarly, corpora are also tagged for text type analysis (Biber, Conrad and Reppen 1998). A third tradition involves the development of corpora for applications such as language learning (as emphasised by Barnbrook 1996) or dictionary-building (in a continuation of the Cobuild project: Sinclair and Renouf 1991) as well as the statistical analysis of texts in authorship studies (Oppenheim 1988). The third approach usually entails an emphasis on statistical properties of the texts rather than parsing procedures. Since I have adopted a view of collocation from Sinclair's and Halliday's perspective, the third approach is particularly relevant to my methods of corpus design and analysis.

The Brown corpus of one million words was one of the first electronic stores of texts for the analysis of English, with the underlying aim to be as representative of the general language as possible (Kučera and Francis 1967). The London-Oslo-Bergen corpus (LOB, Svartvik and Quirk 1980, Svartvik 1992a/b, Leech 1991) was also built up to one million words and was one of the first to attempt coverage of different language varieties, including 15 types of written text – although the texts were artificially curtailed, with a maximum length of 2 000 words. Nevertheless, LOB constituted for some time a major source of data for the study of text types (Biber 1986 *et seq.*). While the first generation of corpora were developed for general linguistic description, the second generation aimed at maximum coverage of the language for the purposes of dictionary-building. These included, in the UK, Birmingham's *Bank of English* (once known as the Cobuild corpus: Sinclair 1991) and the *British National Corpus* (BNC) of Oxford University and Longman (Burnard 1992). These corpora quickly built up the number of texts to hundreds of millions of words by accessing the electronic press and other networks that became available in the early 1990s. Although both corpora

had at one point over two billion words (Sinclair 1993a, Rundell and Stock 1992), each corpus has recently been limited to a selection of just over 100 million words. Another notable corpus project, the Cambridge Language Survey, attempted to build up corpora and develop software in order to compare seven major languages with particular emphasis on developing agreed codings (tags) for semantic, functional and syntactic categories (Atkins, Clear and Ostler 1992). These lexicographic corpora have now been joined by a third generation of more fragmented text collections, including dialect corpora, spoken corpora, restricted language corpora and other specialist text collections (Svartvik 1992:12, Biber, Conrad and Reppen 1998).

As corpora grow in size and complexity, 'representativeness' or an idea of what proportion of texts should be included in the corpus has proven to be a major stumbling block. In his comparison of three major English language corpora (Brown, LOB, and Cobuild), Ljung (1991) points out that within the most frequent 1 000 items of each corpus, 204 words are not shared. Such differences seem to undermine the claims of the corpus-builders that their corpora are representative of the language in general. Ljung further notes very important genre differences between the corpora, especially Cobuild, with its large number of high frequency abstract nouns dealing with domains of behaviour, geometric shape and politics - the kinds of lexical preoccupations to be found in journalism (1991:249). Because of the wide availability of journalistic texts in the initial years of corpus analysis, linguists pointed out that the data in large corpora were susceptible to stylistic bias (Rundell and Stock 1992). While quantitative representation is a problem, there are also artificial barriers to inclusion which arbitrarily restrict the nature of the corpus. For example, Burnard (1992) noted that his own corpus, the BNC had a no-translations policy which eliminated such influential texts as the Bible. Similarly, Collins and Peters (1988) have questioned the motivation behind the text categories of several corpora. They note for instance that LOB gives as much weighting to *belles lettres, biographies* and *essays* as to *the Press* or *learned and scientific writings*.

Nevertheless, genres are by their very nature unequal, and it is perhaps unreasonable to describe the whole language on the basis of equally represented text-types. One might argue that the spoken language and dialogue should make up the vast majority of any general language corpus, since the corpus may wish to represent exposure (from an individual's point of view) rather than textual variety. The other possibility is that each recognised register or genre should have an equal footing because the language system is not wholly represented in the more frequently encountered varieties. These are clearly fundamental questions but with very

few straightforward solutions. It is for this reason that it may be prudent not to scale down the corpus, but to favour the analysis of specialised genres. However, as noted above in terms of the discourse community, even the question of representativeness of a single subject matter (cancer research) appears to be a complex issue.

4. Corpus Analysis and Languages for Specific Purposes

Whereas corpus linguistics has tended to favour the construction of large scale text collections for the analysis of the 'general language', much less work has been carried out on corpora of specific language varieties. McEnery and Wilson (1996) mention that there has been some work on spoken and written variation, but very little work on specific text types. General corpora tend to include sections of technical texts for comparative purposes, but understandably these have been very broad in scope, largely because it has been felt necessary to collect a broad range of subject specialisms. Nevertheless, in the field of *English for Specific Purposes* as mentioned above in Chapter One, a number of linguists have carried out studies on very specific corpora, including Myers (1989), Kretzenbacher (1990), Banks (1994a), Salager-Meyer (1992), Williams (1996), Dubois (1997) and Biber, Conrad and Reppen (1998). A small number of studies have so far dealt with grammatical collocation and genre analysis (Gledhill 1995a and 1995b), or systematic analysis of clusters of grammatical features in technical texts (especially Biber, Conrad and Reppen 1998).

There are a number of studies which have specifically targeted collocations in science within the field of terminology (Thomas 1993, Baker, Francis and Tognini-Bonelli 1993, Pearson 1998). These studies follow the tradition in terminology which distinguishes between collocations in the general language and those in the LSP, a notion which is widespread but which has also been widely criticised (Bloor and Bloor 1985). The position is summed up tersely by Sager et al. (1980:231): the potential for collocation in the general language is freer than in the special language. Benson et al. (1986) have been the principal proponents of this view and have argued that LGP and LSP collocations can be distinguished in terms of their syntactic behaviour. For example, in compound nominals in the LGP, head nouns become more specific as in *cabinet reshuffle* and *drug pusher* and the attributive nature of the second element can be reinforced by reformulating with *'of'* or other grammatical items: *a reshuffle of the cabinet*, *a pusher of drugs, a booster for brakes*. However in LSP compounds such as *measles vaccine, jet engine, house arrest* such deconstruction is not possible. He

claims that LSP nominal groups must have a generic-specific internal structure that distinguishes them from their LGP counterparts (moving from specific to generic). The lack of reformulative potential of a multiword term therefore suggests a systematic means of distinguishing between fixed LSP terms and looser LGP phrases. However, this type of distinction reinforces the traditional view that the LSP is merely a series of grammatical restrictions, and seems to arbitrarily assign LSP or LGP status to items which may have very different distributions (for most observers *brake booster* appears to be an LSP item, regardless of grammatical mutability).

Thomas (1993) provides a more text-based account of LSP phraseology when she describes the types of collocation that occur in a computer based terminological term bank. She finds that in the search for collocational nodes to prioritise as dictionary entries, LSP phrases may use similar resources to the LGP but their predictive collocational elements vary in position from the LGP as the expression moves from left to right. Thomas notes that collocational variability, where the node word is highly predictive of the left or right collocate, affects the lexicographer's choice of base word. Sinclair similarly refers to this phenomenon as a statistical problem of 'up or down direction of collocability' (1987c:330). Contrary to the impression that LSP style is 'highly nominal', Thomas notes that LSP verb phrases have a 'high range of functions and occurrence' including transitives (occlude, induce), intransitives (phase-separate, hydrogen-bond), phrasal intransitives (denatures into, localises in) and are particularly prevalent in passive phrases (is synthesised in, are conserved) (1993:60). More generally, frequent verbs in the LGP become highly predictive of object nouns in the LSP (*to boot a computer, to create a file*) (1993:55). Sager et al. similarly note that the collocability of verbs is limited to phrasal units while nominal groups have taken over the function of representing mental categories, conceptual phenomena and operations (1980:86). They note a tendency for grammatical themes or subjects and descriptive predicates, and the predominant pattern of noun + [copula] + Property / *of* + Property (material - shape - design) (1980:188). They also note inversion in declarative sentences where a past participle (such as *Attached to the X is a Y...*) introduces elements at the thematic beginning of the sentence.

Despite the rarity of corpus work on scientific texts, linguists and stylisticians have identified a vast range of grammatical and lexical properties of virtually every imaginable variety of language. Muller (1968, 1977) has notably established a well known methodology of word counts to establish different authors' styles (Oppenheim 1988, Potter 1990:411). Among corpus analyses of style, Johansson (1982) reports on the untagged analysis of four types of writing from the LOB corpus where he analyses the relative

frequency of function words. Fox (1993) has analysed the frequency of *then* following sentence subjects as a characteristic of the language of law enforcement. Choueka et al. (1983) studied collocation in the language of the New York Times. Butler (1993) studied discontinuous collocational frameworks in Spanish magazines and found that prose articles can be shown to be different to interviews. He found that frameworks contain more textual information in the former and interpersonal, discursive phrases in the latter. Finally, Collot (1991) has examined the use of comparative constructions in e-mail communication. As noted above, with some exceptions (Butler 1993, Banks 1994b, Gledhill 1996 and Williams 1996) the focus of work even in such a large area as stylistics or register studies has been on grammatical categories rather than on collocation and phraseology.

5. The Status of Corpus Evidence

In this section I examine the philosophy underlying different approaches to corpus data, in particular in relation to the notion of item selection (which lexical or grammatical features to identify) and item identification (the use of tags or other methods).

As can be seen from our discussion of the idiom principle, Sinclair and his colleagues assume that there should be as little human involvement as possible in the construction and analysis of a corpus. All grammatical evidence should come from real examples analysed as automatically as possible as opposed to invented ones analysed introspectively. Sinclair distinguishes in this respect between the natural but untidy feel of examples taken from a corpus with the grammatical but odd nature of examples used in theoretical grammars. Although controversial, his main point has been conceded by many generative linguists, who now use corpora if not to elicit data, then at least to check their hypotheses (Blackwell 1987, McEnery and Wilson 1996). The principal research method of the Cobuild research group of the 1980s (Sinclair 1981 *et seq*., Fox 1993, Francis 1985, Clear 1987, Krishnamurthy 1987, Renouf 1987a/b, Hunston and Francis 1998) and researchers who were influenced by the approach (Miall 1992, Gledhill 1996 as well as workers in Cobuild's successor project, the Bank of English) has been to eschew the traditional categories of linguistic analysis to the point where they analyse raw data that has had no prior linguistic treatment (or 'tagging'). On the other hand, many corpus linguists (Leech and Fligelstone 1992, Garside, Leech and Sampson 1987, McEnery and Wilson 1996) are involved in work that changes the format of the texts that they are working with, whether it is by transcribing prosodic markers from spoken texts or by

implementing automatic tagging (marking of word class and syntactic function).

Although Sinclair's 'statistical / textual' view of collocation has been influential, it is not generally accepted by corpus workers outside the Firthian or Hallidayan tradition. Unlike the main thrust of Sinclair's work, the majority of corpus research is conducted on tagged or marked-up corpora, and can benefit from the use of pre-defined categories. A search by a parser or a tagged corpus analyser can be initiated by asking the computer for 'nominals followed by conjunctions' (category tags) or 'indirect complements' (functional tags) and so on. In other words, whereas Sinclair's approach has been to see collocation in all recurrent lexical forms, others limit the kind of expressions that the computer counts as acceptable. The tagging approach is instead used in systems for automated parsing or collocation retrieval (in terminology, the information sciences and in abstracting), where the need to cut down on combinatorial possibilities is considerable (Sparck-Jones 1971, Choueka et al. 1983, Frohman 1990, Ahmad et al. 1991, Bazelli, Pazienza and Velardi 1992, Busch 1992).

Tagged corpora are also used widely to test the hypotheses of formal and generative grammars (McEnery and Wilson 1996 provide an overview of these studies). These approaches traditionally privilege the 'semantic / syntactic' view of collocation I proposed above, largely because they use data to confirm rather than to define instances of collocation. A typical study begins with a definition of collocational relation between words using a lexicalist model and as proceeds to classify any fixed expressions within the framework of that model (for example, Mel'čuk 1988, Fontenelle 1994). Furthermore, since idiomaticity is seen as a structural or formal functional problem within the generative framework, corpus data have also been used to demonstrate the typical grammatical profile of fixed expressions (Ringle 1982, Abeillé 1995). In these studies the fixed expressions are taken as 'given' and derived from existing studies on idiomaticity. Annotated corpora can of course be used to capture the kinds of combination that Sinclair is interested in, but they generally tend to rely on an automatic parser which has already divided and marked the text up into syntactic categories and functions. As a consequence, these approaches conceive of collocations between existing grammatical classes or functions (for example: noun + verb) and do not therefore initiate searches for the kinds of grammatical collocation identified by Sinclair and his colleagues (discussed above).

As an example of a collocation-retrieval approach, Smadja (1993,199) has implemented a program that initially finds collocations on a statistical basis and then uses a 'syntactic filter' to eliminate non-phrases. He tests the results

of the automatic system against four generally-accepted principles of collocation:

Principle 1 Collocations are arbitrary (1993:146).

Collocations are combined as a lexical choice which may not have any semantic or syntactic explanation. This can be seen between languages, where word-to-word translations have different distributions. (*enfoncer la porte* - to break down the door, *enfoncer un clou* - to hammer a nail in).

Principle 2 Collocations are domain-dependent (1993:146)

Collocations have a very specific distribution in terms of technical jargon and terminology.

Principle 3 Collocations are recurrent (1993:147)

Collocations can be accounted for statistically, that is they are not accidents of occurrence or independent variables and are established as a recognisable part of the language (a point also made by Church and Hanks 1989).

Principle 4 Collocations are cohesive lexical clusters (1993:147)

Collocations are internally consistent with elements which are predictive of others. Although this is unlike Halliday's textual definition of cohesion, there is a sense of unity and 'texture' that Halliday and Hasan (1976) refer to within collocations such as *heavy trading,* or *agree to.*

Smadja (1993) suggests that at present his system is good at identifying 'small' collocations (especially phrases which conform to Principles 3 and 4). The types of collocation that Smadja's system is able to identify are listed below:

Type 1 Predictive collocation.

In this type of collocation, one or more elements in the phrase may predict the others, but not necessarily the other way round (*make* and *decision* for example). These collocations are usually flexible in that they may undergo transformations or reformulation without disturbing basic meaning (Smadja 1993:399) and correspond to Cowie's (1981) and Benson's (1989) restricted collocations.

Type 2 Rigid noun phrases.

These are 'important concepts in a domain.' (Smadja 1993:148) such as *stock market* and *Dow Jones* and have been previously studied by Choueka et

al. (1983) in their study of the New York Times corpus and by Burnard (1992:15) who terms them 'text-oriented' co-occurrences.

Type 3 Phrasal templates.

These are collocations which include very free elements within a restricted structure (such as *Stockmarket [X] rose / was up / fell [number] (points) to / at [number]*). These correspond to Renouf and Sinclair's (1991) collocational frameworks and Nattinger and DeCarrico's (1982) phrasal constraints.

Smadja (1996) claims an identification rate of around 70%. While apparently successful, this means that 30% of the terms identified by the *Xtract* system are not valid collocations. The essential problem here is that analysts such as Smadja pre-define a collocation as a valid grammatical phrase, whereas Sinclair and others are prepared to accept collocations which are not constituents of the same phrase or even the same clause. Another difficulty with Smadja's approach is the concept of the non-phrase and the means by which it is possible or desirable to eliminate combinations encountered in the corpus. Non-phrases according to Smadja (1993/1996) are combinations which can not be analysed by a parser and theory identification is therefore dependent on the quality of the parser rather than the quality of the initial data.

From a statistical perspective, Kjellmer (1984:163) has also argued that restrictions are necessary because statistical analysis may throw up either randomly recurrent word combinations (hence *although he, hall to* may occur but are not acceptable phrases) or unusual grammatically restricted sequences (*green ideas, yesterday's evening*). He claims that valid phraseological units are only to be found at the intersection of the two (*last night, try to*). However, Kjellmer (1990) gives much more scope to grammatical collocation than other linguists working on tagged corpora. For example, he finds evidence to suggest that certain grammatical classes are more productive in collocation. Articles and prepositions are involved in the greatest relative number of collocations although their collocates are hard to predict. Singular and mass nouns are similarly highly collocational, but are more predictable in that they have very strong patterns immediately before function words and tend to be premodified in limited ways (1990:167). In addition, verbs have the highest rate of co-occurrence with closed-class items, indicating the important role of phrasal verbs in English, a point also noted by the Cobuild group (Krishnamurthy 1987). These findings are commensurate with many of Sinclair's findings. They also serve to show however that the 'statistical / textual' approach is an ideal, and much work

being carried out from Sinclair's perspective does in fact exploit tagged corpora.

Perhaps one of the more hotly contested points has been over the extent to which it is necessary to mark up the corpus grammatically. The 'collocationalists' and followers of Sinclair argue that since they do not impose traditional grammatical categories, only their approach can achieve original insights about language:

> If [...] the objective is to observe and record behaviour and make generalisations based on observations, a means of recording structures must be devised which depends as little as possible on theory. The more superficial, the better. (Sinclair 1987b:107)

Conversely, Leech and Fligelstone (1992) and others consider that the counting of concordance items is at best 'a trivial facility' and that the only significant data can come from annotated corpora. Aarts is of the opinion that without some degree of syntactic classification, a corpus is useless:

> [...] as everyone knows, the comparison of corpora containing just raw text cannot go beyond linguistically rather trivial observations. (1992:180)

Several corpus linguists have debated the relative success of automatic parsing and tagging (Brekke 1991). Souter (1990) calculated the range and distribution of 8522 syntactic structures found by a 'componence parser' (componence rules are syntactic and functional phrase structure algorithms: such as Subject_NGP_det head). He found that just over 70% of these rules are used only once in his corpus. He concludes that if these results were projected to an even bigger corpus, 'a comprehensive grammar for English could be as open-ended as its vocabulary.' (1990:194). On the other hand, Briscoe (1990) has dismissed this kind of argument. He claims that although 'all grammars leak slightly', there is no evidence for a group of deviant or unique grammatical constructs, arguing that the existence of even large numbers of unique grammatical constructs does not invalidate the applicability of a general underlying generative syntactic principle. Conversely, Church and Mercer (1993:4) state that parsers are useful for understanding 'who did what to whom', but are less useful for predicting likely usage in authentic language. The other disadvantage of parsers is that they have, according to Church and Mercer 'little success in word class or word sense disambiguation' (1993:9).

The benefits of tagging and parsing can not be dismissed lightly. Clearly any system which categorises linguistic evidence would benefit from a computational way of counting and sorting the data (McEnery and Wilson

1996, Barnbrook 1996). In this light, some tag sets have attempted to incorporate 'discourse items'. Svartvik (1993:24) has proposed a 170 tag system with labels such as *greeting, fluency device, hedge* and so on. Linguists who impose tags on a text in such a 'manual' fashion are faced with the difficult task of lemmatisation, whether to treat forms such as *be, is, are* as one or different word types. Lemmatisation is particularly criticised by Sinclair (1991) and Francis (1993) who point out that it is a redundant process because collocational patterns tend to reveal differences between word types: the collocations of *be* are different to the collocations of *is* and this distinction is effectively eliminated if both are counted as the same lexical item. There is also some statistical evidence in support of this. Youmans, in his analysis of the 'velocity' or rate of change of frequency of new words in texts found that lemmatisation does not significantly change the curves of type / token ratios (1991:766). Whatever the accuracy of tagging and parsing, I hope to demonstrate below that the quality of analysis relies just as much on the depth of preparation of material as on the formulae used to arrive at automatic analysis.

The fact remains that manual analysis of unrefined concordances can still reveal much interesting data. This is especially true of features of discourse which do not have categorical forms (such as evaluation, modality, grammatical metaphor, discourse anaphora and so on.) as the work of Stubbs (1996) and others has demonstrated.

One of the more fundamental debates that have been conducted in corpus linguistics centres on Sinclair's claim that corpus work must attempt to account for the naturalness of authentic data rather than a theoretical search for an abstract notion of grammaticality. However, many linguists warn against seeing the corpus as a guarantee of truly objective data. In Fillmore's (1992) analysis of the use of the word *risk* he demonstrates that the word has a unique lexico-grammar in the language in that *'running a risk'* conceptualises harm as a result of an action, while *'taking a risk'* sees harm as a result of a goal. But he cannot see how a computer could ever come to determine such a pattern, or how it could rule out alternative expressions. Chafe takes a similar stance:

> A corpus cannot tell us what is not possible... Should it ever come about that linguistics can be carried out without the intervention and suffering of a native-speaker, I will probably lose interest in the enterprise. (Chafe 1992:59)

In a sense, this argument could be turned around against tagging, since Chafe and Fillmore are discussing linguistic features that appear to be beyond

automatic parsing, but are not beyond more basic empirical quantification. In any case, Chafe, Fillmore and others claim that Sinclair has missed the point about intuition, and has ruled out the important function of negative data in constructing a model of syntactic principles. For them it is important for the model to be able to explain why certain features of language do not occur, and the corpus does not provide this explanatory adequacy. They also point out that there is nothing inauthentic about a native-speaker's intuitions about examples and counter examples (although as we have seen, other generativists have made much use of corpora to test their hypotheses for positive data).

Chafe's point essentially contrasts the generative linguist's preoccupation with selected counter-examples with the empirical linguist's interest in authentically occurring data which is often more difficult to analyse. Sinclair's approach is not concerned with grammaticality but with an account of naturalness in language. Native intuition and invented examples may be enough to explain the underlying syntactic principles of potential expression, but they are inappropriate when we need to address issues of style and textual acceptability. He argues that although the corpus replaces introspection in linguistic analysis (essentially guessing at data and inventing examples), the computer still implies the use of human intuition (a native speaker interpretation, a linguist's skill in explanation), a factor that Fillmore and Chafe appear to have overlooked.

In addition, a corpus of authentic texts is undoubtedly the product of a human intuition, but the linguistic behaviour used to produce authentic texts is uninhibited, unselfconscious and natural. The same can not be said for invented examples or examples created to prove some grammatical point. Sinclair cites a continuum of examples from cryptical to explicit: *we searched* (most cryptical)*, we searched all night, we searched all night for the missing climbers* (most explicit) (1984:206). He asks at what point or in what context each of these kinds of utterance would be deemed to be natural, and suggests that most authentic text occurs at some point in-between. In natural speech, therefore, there is a happy medium between the cryptical and the overtly explicit. This argument for authentic examples has been particularly relevant in the field of lexicography, where the examples chosen for each entry in his Cobuild dictionary were not designed for lexicographic purposes but taken from authentic texts. Furthermore, Sinclair claims that the internal grammatical relations of the sentence are not relevant when one attempts to take account of the function or natural feel of the sentence in context. As with Hoey's discussion of lexical cohesion, we can see how Sinclair's approach moves our attention away from words in a sentence-based grammar to items with a definite textual function.

The discussion in this present section has concentrated on the quality of data analysed in corpus linguistics. I conclude that a tagged corpus and a syntactic parser are not immediately necessary for an analysis of typical corpus style, and note that such processes may indeed by inappropriate for a genre analysis of the type I envisage, at least at the present time. Since my primary aim is to establish a general phraseology of research articles, I hope to show below that instances of collocation can be fruitfully identified on the initial basis of statistical analysis rather than resorting to formulae and syntactic parsing of the sort proposed by Smadja and others.

In the preceding sections, I have set out Halliday and Sinclair's perspectives on discourse analysis and corpus linguistics. Halliday establishes the notion of register as probable expression, and emphasises the changing role of linguistic features as they are used in different rhetorical contexts. In addition, we have seen that Halliday and Sinclair's view of the lexico-grammar prioritises the role of grammatical collocation and grammatical items, and my corpus analysis below therefore concentrates on the phraseology of these items and their distribution within the corpus. The following sections discuss the main steps involved in the corpus analysis and attempt to implement the 'statistical / textual' analysis of the corpus as a first stage in the phraseological analysis of the research article genre.

6. The Corpus and the Discourse Community

> A corpus is a text assembled according to explicit design criteria for a specific purpose, and therefore the rich variety of corpora reflects the diversity of their designers' objectives. (Atkins, Clear and Ostler 1992:13)

It is now necessary to set out the principles underlying my choice of texts for the **Pharmaceutical Sciences Corpus** (PSC). In brief, the PSC contains:

- 150 research articles from 22 different journals on cancer research and pharmacology.
- 500 000 words of text, excluding reference sections, tables and footnotes.

I propose to analyse these texts in terms of their different subsections (Titles, Abstracts, Introductions, and so on) and conduct the analysis by examining the collocations associated with those grammatical items which have been found to be statistically significant within each section.

I have suggested above that corpus analysis presents considerable methodological advantages for a description of languages for specific

purposes. In the first instance, the rhetorical aims of the writers are known and can be prioritised in the analysis: this is not an anonymous collection of texts. In addition, we have seen that while there are many studies of phraseology and lexico-grammar in the general language, few specialist varieties have benefited from a large-scale corpus analysis of this kind. The corpus does not represent the register of science writing, but instead focuses on one genre (the research article) dealing with one very specific discourse (cancer research). The usual problem of representativeness is therefore minimised, although not entirely eliminated.

We have seen above that, historically speaking, corpus projects have tended to opt to represent an entire register or language variety. These projects have often found it difficult to delimit boundaries for their constituent texts. For example, Renouf (1987b) states that the texts used in the Cobuild corpus range from very broad registers (non-fiction, procedures, argument-positional texts and narrative) to very specific genres (surveys, the NATO-corpus, the *Sizewell enquiry* corpus). Since such a disparate collection of texts is not clearly defined, Sinclair (1993), Atkins, Clear and Ostler (1992), Ahmad et al. (1991) and others have argued for a more systematic approach to text types in corpus linguistics. Sinclair (1993c:6-7) proposes four principles of corpus design which I adopt in the following sections:

1. The choice of texts should be governed by a stated view of language in communication.
2. The variables determining the choice of texts should be distinct and identified.
3. The component texts should be clearly identified, described and documented.
4. The proportions of different text types should be clearly stated and concomitant with principle 1).

6.1 The Language View of the Pharmaceutical Sciences Corpus

As stated earlier, the research article – despite its variety of forms - is seen as a privileged statement of public research and is thus a major object of enquiry in linguistics. Other texts, such as grant proposals and internal documents mentioned in my survey can be ruled out of the corpus because they form part of the non-public world of Auger's (1989) 'grey literature'. Instead of exact representation of genres in the discourse community therefore, a rhetorical overview of the department should emerge from a mixture of authors' own research articles. These texts are considered to be central to the researchers' work, and appear in the journals which the researchers regularly use for 'indexical' purposes in the lab and for general research reading.

6.2 Design Criteria of the Corpus.

One cause of imbalance in this and perhaps many other corpora lies in the range of potential criteria for the selection of texts as can be seen below (from Sinclair 1993c: 6-7):

Medium-oriented choice:

1-*Author*	Texts selected from informants' own publications.
2-*Access*	Texts chosen on the basis of free access, machine-readability, etc.

Research-oriented choice:

3-*Journal*	Texts from the same journals as informants' papers.
4-*Prestige*	Texts from recognised or prestige journals.

Topic-oriented choice:

5-*Sample*	Texts from a wide sample of journals which cover the area generally.
6-*Centrality*	Texts or journals considered essential by informants.
7-*Field*	Texts covering one research activity or concern only, perhaps on the basis of bibliography or keywords.
8-*Coverage*	Texts chosen at the level of overview or specialisation.

A combination of these criteria were used to select the texts for the PSC, although some criteria account for more research articles in the corpus than others (especially *author, prestige* and *centrality* but also *access*: see below). Such variables cannot be made entirely distinct. As we saw in the survey of the Pharmaceutical Sciences Department, the fourteen researchers had published in their respective fields, and some of their articles provided a substantial basis for the corpus as a sample of their output. However, their contributions alone would result in a very heterogeneous body of texts, not only in terms of different sub-fields as mentioned above, but in the degree of coverage of the field. For example, one researcher donated an introductory paper taking a long-term view of his work, in a journal which would have had a wide readership: *Trends in Pharmaceutical Sciences* (TPS); whereas another donated an article in the specialised *Tetrahedron Letters* (TL) which was an incomplete part of a series of communications on a specialised drug. Clearly, the readership of such a paper would be highly limited.

In an attempt to collect a representative spread of research articles, one might calibrate the papers by criteria such as 'field', 'centrality' as suggested above, or by classifying journals by 'coverage of subject' (general or

specific) or 'size of expected audience'. Another solution would be to use a measure of prestige. As I mentioned earlier, the department judged its own research publications according to Impact Factor scores. While papers in research selectivity exercises are judged according to a researcher's publications in high-ranking journals (calculated from citations in other journals), the head of the department (*PL*) pointed out that some prestigious and well known journals were misrepresented in the listings. He pointed out that the *Journal of General Microbiology*, a journal subscribed to by the department and mentioned even by chemists in the survey, does not appear in the first 600 journals of the Science Citation Index. It was also noted that the well known high-circulation journal *Nature* (14th position) was at one point preceded by the esoteric *Advanced Cyclic Nucleic Proteins* (8th position) (SCI 1993:83). One explanation of this is that while *Nature* is a widely distributed publication, citations in 'working' journals, perhaps used more indexically than for browsing, are likely to make use of more specific data from less well-known publications. It may therefore be misleading to state that a corpus represents 'prestigious journals in the field', where even an objective measure attempts to distinguish this. Nevertheless, this rather idiosyncratic measure does have some importance, since it is valued by the institution and external funding councils, if not by the individual scientists themselves.

The reputation of journals is also rather difficult to gauge. *Tetrahedron Letters* was of doubtful quality according to another researcher (*DP*), because it published 'accelerated' communications which have not had time to be tested. Others saw it as an important journal for new research. One way around this problem was to ask the scientists to cite specifically the last five papers they had been using as reference material or in the lab and in their periods of writing up. This ensured that the corpus included a wide range of journals and topics.

6.3 Choice of Material in the Corpus

The compilation of the PSC involved 150 research articles from a selection of 22 journals. A full list of these articles and the source journal are set out in Appendix 2. A target of 500 000 words was set as the initial corpus size. In order to reach this target after the initial collection of papers from the authors in the survey (which gave 46 papers, criteria 1 and 2, below), a further 104 random papers were selected according to prestige and accessibility (criteria 3 and 4, below). The number of articles collected from each journal was largely determined by how many papers could be copied a factor limited by copyright restrictions (usually one paper from each issue was permitted for

research purposes). But equally crucial were the length of the article and quality of paper for scanning. The following conditions of inclusion in the corpus emerged:

1- *Authorial*: The corpus includes 10 research articles authored or co-authored by interviewees. One researcher submitted three papers, another two papers (one in electronic form) and five others submitted one each (one in electronic form). Four researchers did not donate an article.

2- *Centrality*: The corpus includes research articles from journals mentioned in survey question 5b (specific papers the researchers had recently). 36 articles were obtained in this way, mainly from the ADONIS biochemistry on-line catalogue.

3- *Prestige*: The corpus includes 80 research articles from journals mentioned more than twice in survey question 5a (journals the researches considered important in their field, but which they had not necessarily consulted recently).

4- Accessibility: The journals FAT, JPP and CAR were available on Medline and could
be immediately downloaded (abbreviations refer to journal titles listed in Appendix 2). Article AC was submitted by a researcher from Birmingham University. This gave 24 articles.

In Appendix 2 the corpus is documented in terms of Journal SCI Rank, percentage size of the corpus per journal and title of each research article. The topical and textual breakdown of the texts are detailed in section 6.6.

Choice of Articles and Numbers of Papers.

1.	By author:	BJ, CC, JCPT[7, 8, 9, 10], JMC, JNCI, TL, TPS
2.	By topic centrality:	BJC[1-11], CL[1-9], JGM[1-9], JOC[1-7]
3.	By prestige:	BJP[1-3], BMJ[1-5], CCP[1-16], CR[1-12], IJC[1-25], JCPT[1-6], JOACS[1-11], PAH[1-2]
4.	By accessibility:	AC, CAR[1-10], FAT[1-10], JPP[1-3]

It was decided that the PSC would be split into several subcorpora (pharmacology and cancer – the main division within the pharmaceutical sciences department) but also into sections including Titles and Abstracts (as subgenres in the research article) and Introduction, Methods, Results and Discussion subsections (TAIMRD). Although the original 150 Titles and

Abstracts of the PSC are compared directly with other rhetorical sections, an additional subcorpus was deemed to be necessary in order to obtain more results. This was derived from the electronic index *Medline*. The *PSC-Medline* subcorpus consists of the first 572 abstracts (58 332 running words) selected by the keyword 'cancer' in December 1993. The subcorpus also includes a separate text of the 572 corresponding Titles (7 626 tokens) for comparison with the Abstracts. The Abstracts are all author-abstracts, from a very wide variety of English-language journals and relate to cancer either from within the Title or Abstract or from the list of keywords included as *Medline* data (the keywords are discarded for this study). The Medline corpus thus has the advantage of topical specificity as well as being a homogenous source of scientific texts. In the data analysis section, I compare the PSC titles subcorpus with the PSC as a whole to give a picture of the salient lexical items which are typical of titles with the PSC. These results can then be analysed using the Medline corpus, since the PSC titles corpus alone is not large enough to reveal interesting concordance data.

A number of scanning mistakes due to small print account for certain anomalies of word counts in my data. In many cases, this meant that some experimental sections had to be discarded as they often have smaller print than the rest of the article. The texts that accompany tables were also eliminated unless they had a considerable amount of argumentation, in which case they were considered to be valuable parts of the rhetorical section in which they were situated and added to the end of that section. Once post-edited, all the texts were converted to text files for use on a PC mounted UNIX system for frequency tests and then converted to text files for analysis by a PC wordlist and concordance package (detailed below).

The PSC thus consists of 150 research articles, consisting on average of 7 sections each. Using Roe's word analysis programs (1993b:10) a UNIX word frequency count calculates the total word count to be 515 073 running words (tokens) (Roe takes a word to consist of any string of symbols bound by two spaces, excluding figures). However, this number of words is probably too large (some chemical symbols, Greek letters and mis-scans are also identified by this procedure). A second count by the Wordlist program (Scott 1993) gave 499 105 words, of which 24 253 were different words (types). The PSC was then split into sections (including Abstracts) and counted using the UNIX wordcount (percentages have been adjusted to take account of overlapping sections such as MR and RD sections):

Table 1: Size of Corpus by Sections.

Subgenre	**(Total)**	**Tokens**	**% of PSC.**
T-Title	(150)	2 123	0.5
A-Abstract	(150)	29 283	6.6
I-Introduction	(150)	60 809	13.7
M-Methods	(125)	113 089	25.5
[MR-Methods/Results	(3)	3 207	(32.0)]
[E-Experimental	(21)	30 759	(47.0)]
R- Results	(120)	123 084	27.8
[RD-Results/Discussion	(27)	37 372	(46.1)]
D-Discussion	(125)	114 205	25.8
[C-Conclusions	(4)	1 022	n/a]
[S-Summary	(1)	120	n/a]
Total (TAIMRD only)		*442 593*	*100%*
[Total (all sections)		513 931	N/a]

In some journals, hybrid rhetorical sections replace the function of two separate sections (Methods/Results, Results/Discussion). For example, the structural chemistry journal JCPT has both RD and E-sections. There are hybrid rhetorical sections in 30 articles as well as nine non-hybrid articles which include additional experimental sections. Nine of the 30 RD-sections are accompanied by experimental sections. Experimental sections occur almost always in chemical and pharmaceutical papers (with the exception of TPS). RD-sections occur mostly in cancer research and microbiology papers. Although these figures suggest they are large sections, they are proportionally smaller than the corresponding non-hybrid sections when these are combined. MR and RD sections are usually indicative of an 'accelerated' publication or communication, especially in microbiology. The relative sizes of the rhetorical sections, as well as an element of overlapping means that statistical comparison between rhetorical sections becomes complicated. Since Experimental sections never replace Methods sections, and are roughly equivalent, these are conflated to M-sections (making the combined section 28.5% of the corpus). It is worth noting here that all Methods, Methods / Results and Experimental sections are combined for the purposes of statistical analysis but Results-Discussion sections are kept separate from the Results and Discussion subcorpora. Results-Discussion sections are taken into account in the statistics for the whole corpus but are not the subject of phraseological analysis in this book. It would be for a future study to determine to what extent phraseology in RD sections is more or less characteristic of R and D sections separately. For our purposes

therefore, we look only at the traditional TAIMRD sections, bearing in mind that an additional control corpus (Medline) is used in conjunction with Titles and Abstracts.

In terms of impact, coverage and prestige (where the latter term simply denotes popularity among the expert informants), the SCI index indicates that some journals in the corpus rate very highly in a list of 8 000 journals, but not necessarily according to the classification obtained from my survey ('prestigious' journals identified by the expert informants are underlined for comparison. 'Prestige' journals have lower rank score):

Table 2. SCI Impact Ratings of the PSC Journals.

Journal Name	**SCI Rank (1988)**	**Journal**	**SCI Rank (1988)**
BJP	84	CAR	326
AC	93	BJC	340
TPS	94	CC	361
JOACS	113	JCPT	370
CR	132	JOC	394
BJ	152	JMC	397
IJC	226	TL	476
BMJ	232	PAH	516

[*JNCI, CCP, CL, FAT, JGM* and *JPP* are not ranked within the first 600]

In terms of relating the PSC with its discourse community, the PSC therefore includes many high impact journals, and has quite a specialised coverage with the exception of such 'introductory' articles as TPS. It is surprising that CCP (*Cancer Chemotherapy and Pharmacology*) is not a 'very high' prestige journal : it was mentioned by researchers from both sides of the department as a key link between them, as the title of the journal suggests.

Having compiled the PSC, the next stage involves a topical overview of the specialisms covered in each research article. Two researchers (one from each main division) helped to classify and gloss all the research articles in the PSC according to the following research categories:

Oncology (Cancer Research Total=83 articles)

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.

The corpus emerges with a large number of papers on the biology of cancer (55% of the PSC), covering a range of probably the most important cancer specialisms, from descriptions of the problem to testing biochemical solutions to the problem (chemotherapy and immunohistochemistry), the latter forming the larger part of the cancer research division. The minority part of the corpus, pharmaceutical sciences (42%) is more diverse, covering more specialisms than is perhaps suggested by the term 'structural chemistry'. As can be seen in Appendix A some journals are topic-specific being mostly pharmaceutical and low impact (BJP, CCP, FAT, JCPT, JOACS, JOC, JPP, PAH) while others have a range of specialisms (BMJ, BJC, CAR, CL, CR, IJC, JGM) and tend to be high impact cancer research / microbiology journals. The *British Journal of Medicine* was one of the most favoured journals, (more than five mentions). Unfortunately, no examples of BMJ papers on cancer were available, so five random papers were included as examples of the genre.

6.4 Corpus Typology

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

As well as considering the internal linguistic features of the corpus, it is necessary to set out systematically the external contextual characteristics of the texts as a whole. As I have already mentioned, one of the more interesting

aspects of corpus design is not an attempt to provide total coverage or representativeness, but the realisation that the texts of even such a specialised corpus are different and distinct. No two corpora can be exactly comparable. With this complexity in mind, Atkins, Clear and Ostler (1992:15-19) set out a taxonomy of corpora for the description of their *International Corpus of English* on the basis of Enkvist's (1989) concept of textual 'context' in corpus linguistics. They propose a typological template to establish the various features of any corpus. In their terms, the PSC can be characterised as follows:

- PSC function is 'informative, persuasive' rather than 'instructional'.
- PSC setting is based on a 'scientific research' setting, including laboratory and institutional use.
- PSC style is 'academic scientific' and presumably varies according to internal factors such as 'technicality' (degree of specialisation).
- PSC technicality is 'high degree of specialist/technical knowledge of the author and target readership/audience'.
- PSC topic is a complex of 'science, biology, chemistry etc.'
- PSC genre is 'research article in the pharmaceutical sciences' but because of varying reader motivations (browsing, reference indexing) and of variations in format and text type (communications, quasi-reports, experimental reports, introductory essays) the term 'research article' covers a wider range of texts than originally conceived. I propose the informal term co-genre for these, and subgenre for such sections as 'Titles', 'Introductions' etc.

It is difficult to establish the other criteria proposed by Atkins et al.. For example, the 'authority' of each text is only known for the texts originating from the survey. Despite the large number of multi-author texts, there is no evidence to suggest that single authorship is indicative of coverage or authority: single-author papers AC and TL are very specific and written by post-doctoral research fellows, CC is a specialist single author text by a senior lecturer, and TPS is a more general text by a professor who also happens to be an editor of other journals. The other factors cited by Atkins et al. can not be easily identified for this corpus. For example, I have no record of the degree of proficiency in English of many of my authors, although many of the co-authored texts appear to be written by scientists from non-English speaking counties.

It is possible of course to analyse any number of these different dimensions from the point of view of phraseology (the phraseology of genetics articles versus structural chemistry, of single-author versus multiple-author texts, or native-author versus non-native author texts etc.). Although

such analysis would be of benefit to the genre analysis of the research article, the rhetorical sub-section of the article remains the main focus of analysis in this book and should serve as a model for future analysis of other dimensions.

6.5 Text Analysis

In this section, I set out the main analytical procedures involved in my analysis of the Pharmaceutical Sciences Corpus. The statistical analysis of the PSC follows the following plan:

1. Frequency: the corpus is split into sub-sections (or 'sub-genres') and wordlists are prepared for each section.
2. Saliency: The *Wordlist* program compares each sub-list with the overall PSC. The most statistically significant grammatical items are selected as typical of each different PSC sub-section.
3. Concordance: The *Microconcord* program is used to establish the collocational patterns of each salient grammatical item. A phraseology for each sub-section can then be established.

The procedure used to prepare and compile the PSC is similar to that used in the compilation of the Cobuild dictionary (as set out by Krishnamurthy 1987, Clear 1987 and Sinclair 1991) and has been broken down into a series of computational steps by Roe (1993a:10-13) on a UNIX-based system called the *ASTEC* suite and later developed for the WINDOWS environment as the *Aston Text Analyser* (ATA). Burnard (1992:21) describes UNIX in terms of libraries of routines used for common procedures that can be integrated into a common environment. While this makes the *ASTEC* analysis extremely flexible, commercially available programs emphasise the presentation of data which is an important consideration in concordance analysis. Further steps in the analysis as well as comparison of the rhetorical sections were thus carried out at a later stage by an PC-based collocation program (*Microconcord*: Johns and Scott 1993) and the wordlist compiler (*Wordlist*: Scott 1993). The differences in definitions of what is an acceptable and unacceptable 'word' in these programs, and textual changes of format in converting the PSC for these systems mean that consequent differences in word frequency lists must be taken into account.

STAGE 1: ANALYSING FREQUENCY. The main justification for using frequency lists in this book is the capacity of the computer to identify statistically the most salient lexical differences between two texts or corpora. We can demonstrate this by preparing a sample comparison of most frequent

words in the PSC with the 17 million word Cobuild corpus (these figures differ slightly from the *Wordlist* generated list in Appendix 1). This is calculated by the ASTEC program by simply comparing two frequency lists as follows:

Table 3: The Astec 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 ASTEC comparison reveals clear differences between the specialist and the general corpora, especially in the sharp increase in the proportion of many prepositions in the PSC (this increase can be more clearly seen in the first 100 words of the PSC in Appendix 1). It is also notable that the conjunction / pronoun *that* at rank 7 in the general language corpus drops to rank 12 in the PSC (with 3 359 occurrences) and the pronoun *it* at rank 8 in Cobuild drops down to rank 41 in PSC (with 1 006 occurrences).

As part of ASTEC, the 'COMMON' program produced a list in descending order of relative frequency of each item in the PSC and a figure indicating the relative frequency in the Cobuild list. A clear pattern emerges from this analysis: clumps of words are very significantly associated with the PSC in the mid-range level of frequency as one would expect (*between, human, table, using, results, both, study, shown, protein, observed, DNA, data* are all at 0.4% or more compared to their occurrence in Cobuild: 0.14% or less). Other higher frequency words have a slightly higher relative frequency in the PSC: *of, and, in, was, with, for, were, by, cell*s, *at, from, or, et al., these, after, also, mice, activity* (all at 0.7% frequency or more in the PSC). Conversely, several grammatical items have a significantly higher percentage frequency in Cobuild than in the PSC*: the, a, to, that, is, as, on, this, are, be, not, which, an, have, it, all, has, but, other*.

Even a cursory glance at these lists suggests considerable differences of grammatical and phraseological patterning between scientific texts and a general language corpus. A number of these differences are examined in more detail below.

STAGE 2: DETERMINING SALIENT WORDS. A **salient word** is a word that occurs significantly more in one text (or part of a text) than it does in another. Using the *Wordlist* program, ten of the most statistically salient grammatical items from each subcorpus were identified in order to examine their collocational properties and phraseology.

The *Wordlist* program create frequency lists and compares them. The resultant 'keyword' list places those words that are more frequent in the text type at the top, and words that are untypical of that text towards the bottom of the list. The first step in saliency analysis involves the *Wordlist* program which compares proportional frequency lists made for each rhetorical section of the corpus, weighing the frequency of words in each list against the proportion of the corpus made up by the subgenre. *Wordlist* then compares the word frequency list of each section with the whole corpus (or part of the corpus if comparing R- and D-sections) providing a chi-square score of significant difference (as described by 1985a and Barnbrook 1996). This is obtained by dividing the observed frequency of the word in the sublist by the observed frequency in the whole PSC and multiplying by the expected frequency, a proportion based on the size of the subcorpus relative to the whole PSC. *Wordlist* then prepares a list of salient words for that rhetorical section. The results of the most statistically significant salient words for each rhetorical section are listed in Appendices 3-8. I have only listed the first 50 items from each result: a *Wordlist* comparison assesses every word including all the words that are non-significant. Unfortunately, these lists are too long to be included in the Appendices.

To demonstrate the use of these saliency lists, here is an extract from the list of salient items in Abstracts:

Table 4: Wordlist : Abstract-salient words in the PSC.

PSC

Rank	Word	Freq. in Abstracts	% in Abstracts	Freq. in PSC	% in PSC	(%)	Chi2	Probability
31	but	67	(0.2%)	663	(0.1%)	18.1		0.000
32	immortalized	13	(0.0%)	69	(0.0%)	17.9		
33	showed	43	(0.1%)	375	(0.0%)	17.4		0.000
34	increased	43	(0.1%)	376	(0.0%)	17.2		0.000
35	interval	12	(0.0%)	56	(0.0%)	16.9		

Items at the top of the word list are relatively more frequent than those near the bottom. This represents the first page of several, so all of these words are particularly 'salient' or typical of Abstracts. Near the bottom of the list in Appendix 4, it can be seen that *immortalized* is the 32nd most Abstract-salient word (by virtue of its observed frequency in the Abstract, i.e. 13 tokens). This result is divided by the observed frequency of the word in the PSC (69 tokens). Its occurrence is not judged by the program to be significant (the chi-square is calculated as 17.9 but a *p* score is not shown). In fact, from the *Wordlist* tables it can be seen that there is a statistical cut-off point in terms of items that are too 'infrequent' compared to items from the whole corpus. For Abstracts the cut-off point is 90. This means that while items with fewer than 90 occurrences in the PSC may be very frequent in Abstracts (i.e. 'salient'), they are not given a p-score.

On the other hand, *but* is the 31st most abstract-salient word, the first grammatical item on the list and has a chi-square score of 18.1, which at 1 degree of difference (Butler 1985a:176) places it even below the 0.1% level. This is considered to be 'highly significant' (5% or less is regarded as 'significant') and those items with a $p = 0.000$ score in the lists are all considered statistically very highly significant. *Wordlist* signals words that are important to the corpus as a whole by showing their percentage if it is greater than 0.1% (in the case of *but* 0.2%). As a statistically salient word as well as a grammatical item, *but* therefore merits out attention. This word is

then listed as the 1st Abstract-salient item in Appendix 4 (followed by *these* and *of* and the other salient grammatical items from Abstracts).

As internal measurements of the relative distribution of words in the corpus, the *Wordlist* results serve as the basis for deciding which items are of interest in our analysis. The assumption here is that a significantly frequent item is likely to play some role in a phraseological pattern. The assumption is also that the significance of an item in one part of the corpus may be typical of that rhetorical section, although clearly an analysis of the use of the word would need to be undertaken across the corpus to rule out overgeneralisation. In theory, a word may have a constant distribution but a different phraseological pattern throughout the corpus. For this reason, those items which have been found to be salient in different sections are analysed in sequence in order to demonstrate any similarities or differences in behaviour.

It is important to note here that chi-squared has recently been criticised for some samples (Clear 1993, Kilgariff 1996) because it compares texts with an idealised notion of general distribution. Kilgariff's observations suggest that two versions of a British English corpus show more variance under chi-square than when American and British corpora are compared. His argument is perfectly reasonable: since no two isolated sentences will share the same distribution of grammatical items, there should be no surprise that high frequency words do in fact vary even within what is supposed to be a homogenous corpus. My argument would be that similar genres have similar grammatical profiles, and that Nevertheless, it should be clear from the Appendices 3-8 that the items identified as salient are indeed very highly significantly more frequent in different subsections of the corpus that one would normally expect in a general distribution (or at least some items are salient in a number of sections, indicating that they are very untypical in others). The ultimate test is that the phraseology which emerges should conform in some respects to previous research which has examined differences in research article subsections, and I signal these instances as necessary in the analysis below.

The subcorpora-salient words that emerge from the *Wordlist* analysis are set out in section IV (data analysis). The rationale for choosing the first ten grammatical items rather than just the first ten salient items in a subcorpus has been discussed above. The main argument is that grammatical items have been relatively neglected in traditional analyses of phraseology, although recent corpus research has emphasised their role in grammatical collocation and collocational frameworks (Gerson 1989). I hope to demonstrate throughout section IV below that grammatical items have very distinctive collocational properties. The significance of grammatical phraseology can be simply illustrated here by fact that the grammatical item *but* identified above

is more likely to be of interest to a discussion of the phraseology of Abstracts than the word *Summary* which is the most salient item in the list, but which is clearly also expected to occur at the head of the Abstract or Summary section! In any case, grammatical items such as *but* tend to be the most salient items in the list (this can be seen in the results for the main sections of the article: Appendices 5-8, although admittedly the results for grammatical items are less striking for the shorter Titles and Abstracts). Nevertheless, many lexical items are also important indicators of phraseology, and I raise any interesting tendencies when I discuss each individual section in Chapter Three. The importance placed on grammatical items here should however not detract from the initial assumptions I have argued throughout this book, that lexical and grammatical items ultimately operate on a continuum.

Some initial results are worth mentioning at this point. The following grammatical items were identified by *Wordlist* as salient words in the different parts of the corpus (I indicate by code the original subcorpus of each item. Some items, like 'both' or 'this' are listed by their most frequent word class as observed in the corpus):

Auxiliary / Modal verbs	(11): was (A, M), did (A, R). been (I), has (I), have (I, D), is (I, D), can (I), were (M), had (R), be (D), may (D).
Prepositions	(11): of (T, A, I), for (T, M), on (T), in (T, A, R, D), to (I), at (M), from (M), after (M, R)
Determiners	(8): these (A), such (I), each (M), no (R), the (R), all (R), our (D), this (D)
Conjunctions	(5): and (T, M), but (A), that (A, D), both (A), when (R)
Pronouns	(4): there (A, R), who (A), it (I), we (I, D),
Grammatical Adverbs	(2): then (M), not (R, D)

The analysis covers 38 items in total, and certain items are salient in a number of different sections of the research article. As mentioned above, this allows for an analysis of phraseological distribution across the corpus: the behaviour of *in* for example, can be analysed in Titles, Abstracts and Results and Discussion sections. The salience of *in* in these sections can be regarded as a result of its relative infrequency of use elsewhere (in Methods and Introductions). Below I set out the analysis in two different ways: by grammatical item (thus examining the changing phraseology of one item

throughout the corpus) and by rhetorical section (establishing a specific phraseology for each sub-section).

STAGE 3: CONCORDANCE ANALYSIS. The first step in recognising patterns in the corpus is to create a computer-readable index of the location of every word in the text, a process that is fully automatic in most concordancing packages.. Patterns of use are made easier to see by placing each instance of a word and its context in the centre of the computer screen (the 'concordance') and changing the list format so that words to the left or the right are presented together and alphabetically. In *Microconcord,* patterns can be calculated statistically (for left, right and total collocates of a word) and the patterns can also be outlined in colour, highlighting patterns over a long range and permitting the analysis and sorting of collocational frameworks (Renouf and Sinclair 1991). Here is an example of an ordered concordance of the word *of* elicited from the *Medline* corpus where the left hand pattern was revealed first; then an ordered listing is elicited for one word to the right:

1 Table 5: Selection from an ordered concordance of <u>of</u>

Anesthetic... management	**of**	a patient with Bartter's syndrome.
neurosurgical... management	**of**	brain {metastasis} from colorectal
Psychological... management	**of**	breast cancer patients in a group.
ort review. 371 Management	**of**	chemotherapy-induced neutropenic
Teicoplanin in the Management	**of**	Febrile Episodes in Neutropenic
Ch resistance in the management	**of**	head and neck cancer.
current trends in the management	**of**	invasive bladder cancer.
current trends in the management	**of**	localised prostate cancer.
irradiation in the management	**of**	patients with liver {metastases}:
{interdisciplinary} management	**of**	...retinoblastoma.
Diagnosis and management	**of**	salivary dysfunction.

From this we can gather that the expression *...management of...* is an important way of introducing the concept of a specific treatment of disease in the title (at least in cancer research). I have imposed a notational convention on the concordances presented in this book as follows:

Bold item	a node word or word currently under

	investigation.
Underlined item	a highly frequent collocate of the node word.
{Item in curly brackets}	a cluster of semantically related lexical items.
<Items in angled brackets>	a fixed sequence of collocates.

We can see from the example concordances that the fixed sequence <in the management of> is not just a phrase in itself but is related to a broader phraseology. This is because it collocates with a consistent set of topical patterns with few deviations from the pattern. For example, the expression is introduced by a general statement of research, in particular the collocations *current trends in, diagnosis and*... or a less fixed and more varied semantic set (clinical histochemical approaches: {*Treicoplanin in, irradiation in, resistance in*...}). However, the word management on its own has a different phraseology. It allows the researcher to signal the general methodology to be undertaken in the rest of the article: {*anesthetic, neurosurgical, psychological, interdisciplinary*}. Similar modification of the type of *cancer* is also involved to the right of the expression and these could be said to be typical processes of inclusion of methodology and precision of problem in the noun phrases of titles.

The advantage of this kind of visual analysis is that it reveals patterns that may not easily be revealed by automatically derived collocation counts. Having identified a pattern such as *management of*, it can be seen that the expression is semantically modified by a topic that is only intuitively accessible: a statement of the disease or its symptoms (*Y cancer*, *Y patients*). The visual cues are not used in all cases, but it can be immediately gathered from the above example that the term *management* involves two consistent phraseologies.

In order to signal where a reading of the concordance has revealed a large scale lexical pattern, a semantic covering term is expressed in brackets and in small capitals {DISEASE Y}. In the phraseological analysis section of the book I have identified four major semantic categories: RESEARCH, CLINICAL, EMPIRICAL and BIOCHEMICAL, with certain further subcategories. I have also used the symbol X to demonstrate the many types of treatment-related names of compounds (often with positive connotations), and Y for many disease-related items. Finally, in order to make the optimum use of examples, a maximum of five concordance lines is usually shown for each pattern.

STAGE 4: CALCULATING COLLOCATION. For my purposes, collocation is a statistical phenomenon of language that can be used to justify the

identification of patterns by the analysis of concordances of a specific context. For example, in the *Medline* control corpus, *management* was found to be not only a frequent but also a significant collocate of *of*. '*Of*' itself was a significant word in titles when compared with the rest of the corpus. Thus the justification of analysis of the initial node **of** and hence expressions in which it plays a role, are based on some comparison with a norm. The term 'statistical collocation' is thus seen as the justification for the assignment of phraseological patterns. The term 'phraseological collocation' is used here to signify patterns that are not significant or even frequent by themselves but are visibly (or intuitively) part of the phraseology, such as the pattern {EMPIRICAL PROCESS} in the management **of** + {DISEASE Y}.

A built-in assumption of statistical collocation (as opposed to phraseological collocation) is that the closer collocates are to their nodes, the greater the collocational force between them. This has led to dispute over the amount of co-text (the span to the left or right of a node) that should be taken into account, on the grounds that, as Sinclair argued, collocates are not independent variables. If so, there should be some systematic approach to determining statistical dependence. Generally, phraseological studies either treat collocation as *directional* (either left of or right of the node) or *informational* (collocates are calculated for both sides). They also vary in the value they assign to the position of the collocate. Thus a different value can be either assigned *locally* for each position of each collocate: first left, second left, first right, second right and so on, or assigned *globally* to a collocate regardless of position or span. Different collocation programs provide a range of means of calculating frequency of collocation (to a span of ten) and position of collocation (to a span of three):

1. *Microconcord:* Short range (3 x 3) globalised collocation (either informational or directional)
2. *Astec:* Short range (3 x 3) localised collocation (directional only)
3. *Wordlist:* Long range (10 x 10) globalised collocation (either informational or directional)

Each of the programs has statistical and analytical advantages and drawbacks. *Astec's SYN* program calculates collocations for all items to the left of the node and the right of the node separately for a span of 3 x 3. Thus the first line for *of* from the PSC is:

the (174)	a (134)	the (574)	**of**	the (354)	of (67)	a (34)

This is useful for determining distribution according to position, but does not give an immediate pattern that can be followed up by closer analysis by

concordance. *Microconcord*, on the other hand, gives equal value to collocates up to a span of 3 x 3. Thus, in the PSC Medline corpus, the first three left collocates of *of* are *the* (100), *and* (59) and *cancer* (41) while right collocates are *the* (78) *cancer* (69) and *in* (63). The program gives at the same time a view of the main concordance and the full co-text, allowing an immediate overview of phraseological patterns in which a word may be involved. *Wordlist* calculates global collocation to a wider span of 10 x 10. The results are more dispersed than those of *Microconcord*, as shown below:

*Table 6: Collocates of '***of***' in a 10 x 10 span, according to the Wordlist program.*

Collocate	*Frequency of left collocation.*	*Frequency of right collocation.*
of	1421	1451
cancer	1203	1295
in	1208	1251
the	1156	1116
a	492	447
with	376	392
breast	279	328
for	359	229
patients	254	258
cell	259	231
human	175	259

This shows that patterns appear to be established even across such a wide span (**of** + *breast*, **of** + *human*). The program also allows for a distribution analysis not across several texts but within a text, giving a 'bar code' of the co-occurrence of up to three items. In his own collocation program, Clear (1993) takes a window of 5 words i.e. a span of 2 x 2 (two words to the left of a node, the node itself, two words to the right of a node) and does not take into account whether items are left or right collocates: they are all calculated together. Clear uses two principles of information retrieval from corpora. *Precision* is the measure of how successfully the system retrieves interesting data. *Recall* is a measure of how much interesting data are actually found and how much are lost. Phillips (1985) and Smadja (1993a) aim at a total collocational description of a corpus, and thus *recall* is an important concept for them. For the purposes of this book, however, *precision* is a sufficient measure of the significance of what Clear terms mutual information.

Atkins, Calzolari and Picchi (1992) define mutual information for collocation as the logarithm (to base 2) of the observed co-occurrence of a collocate with a node divided by the independent probability of either meeting by chance within the corpus. The result is squared to give a steadily increasing logarithmic MI score, where the highest scoring items are considered the most 'collocational'. The following table illustrates the fact that highly mutually informational collocates do not correspond to the most frequent collocates (here the collocations are derived from *Microconcord*):

Table 7: Mutual information (MI) of collocates of the word of from the Medline titles subcorpus.

Collocate	***Corpus Rank***	***Frequency of collocation.***	***MI score. Log P(Obs/Exp)^2***
presentation + of	10	7	8.4
department + of	17	10	8.0
concentration + of	34	17	7.6
majority + of	13	6	7.4
significance + of	24	10	7.2
died + of	28	10	6.8
management + of	43	15	6.8
[...]			
of + patients	11	24	2.0
of + of	2	85	1.7
of + was	9	16	1.4

The MI score also reveals different patterns: it is only until the last half of the MI table for *of* (see the Analysis section 11.1 and Appendix C for full details) that right-hand collocates appear, suggesting that the use of *of* is largely motivated by a limited set of left-hand research-activity or empirically oriented words like *presentation, department, majority, measurement* which are then qualified by a more diverse group of disease-related items (*disease Y, cancer X, patient*...). This example illustrates the fact that frequency and significance only tell half the story: there may be collocational patterns to be discerned in the less statistically salient parts of the table.

For a number of reasons the MI score was not used in the main analysis of this book. To begin with, I examined fifty collocations of ***of*** to obtain the above table. If ten items from each rhetorical section were analysed, I would have to calculate a large number of collocates for each of the 38 items: that means 1900 (38 x 50) two-word combinations. Since I am interested in

longer collocational patterns than 2 words, such an analysis would not be mathematically accurate. This is the reasoning behind Howarth's reticence over automatic identification of phrasemes (1996). Another problem with collocational counts is that some items are significant yet have few short-range collocational properties (such as the statistically significant use of *but* in the abstract). Kaye (1990) suggests that sampling be carried out over a large amount of text to include discussion of long-range collocation such as *so ... as*. In a relatively small corpus such as the PSC, however, most of the occurrences of an item such as **of** can be analysed, since the highest frequency items in the corpus display remarkably stable collocational properties.

To summarise: collocational patterns are identified firstly in terms of raw frequency in this book within a span of 3 x 3 while more diverse patterns are established by concordance analysis. No automatic method (such as the MI score) is applied. Statistical collocations (signalled here by underlining) are therefore a measure of rank occurrence within the span of the node word, but no statistical significance is claimed for phraseological patterns as a whole (in particular involving semantically-related items).

IV. Collocations and the Research Article

The context and specificity of the research article genre have been explored in the introductory sections of this book. A theory of text has been proposed in which collocations and phraseology are seen as central to the discourse of science. In order to examine the research article genre more systematically, the construction of the Pharmaceutical Sciences Corpus (PSC) was described in section III. In this section, I examine the specific phraseological and collocational properties of the corpus with a view to exploring the typical style of scientific texts.

The description throughout the following sections attempts to answer a basic hypothesis about the research article: collocational patterns are assumed to correspond to rhetorical functions, and are also considered to be consistent within different sections of the cancer research article (the so-called rhetorical sections: Title, Abstract, Introduction, Methods, Results and Discussion). In order to examine this specific claim, I set out firstly a separate analysis of those grammatical items of statistical significance in different research article sections (at times this extends to four sections per item). On the basis of the remaining grammatical items (those which are only salient in one specific section), I then examine the particular phraseology of each rhetorical section in turn.

1. Collocations of Salient Words in the Pharmaceutical Sciences Corpus

As explained in section III.6, a *Wordlist* analysis of all the words in a section of the corpus provides us with a systematic comparison of the section and the corpus as a whole. The most statistically significant items are termed salient words (as listed in Appendices 3-8), and these items can be sorted according to three criteria:

1. significant lexical items.
2. significant items of high frequency in the PSC.
3. significant grammatical items.

In my discussion of data collection above, I argued that grammatical items give the optimum amount of phraseological information for a medium-to-small sized corpus such as the PSC. As we have seen, statistically the PSC is too small to provide interesting phraseological data for low frequency items (criterion 1) and in such cases *Wordlist* imposes a statistically-determined cut-off for each section (those items which do not obtain a *p=000* score). It can be seen that many such criterion-1 items are very specific lexical items or *hapax legomena* (accidents or very or unique forms such as B6C3F1 in the Title-salient list). Criterion 2 on the other hand provides an immense amount of valid data, as can be seen in the results for Titles and Abstracts (Appendices 3 and 4). My argument for criterion 3 simply rests on the assumption that an analysis of phraseology from the basis of grammatical items minimises the amount of data analysis needed by characterising global patterns first. I maintain that the kind of data obtained under criterion 2 would be more suitable for a lexicographic or terminological survey than a phraseological one. As we have seen, 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 often too large to provide insights about specific text-types. And it has been shown in our discussion of the lexico-grammar that many phraseological units contain at least one grammatical item. In other words, if grammatical items are analysed as a priority over and above criterion 2 items, then it follows that lexical items of interest should emerge as organising elements within a larger phraseology. In most cases, as can be seen in Appendices 5-8, grammatical items are more frequent in any case, and it is likely that any patterns they display will be more statistically significant than those of lower frequency lexical items.

As detailed in section III.6 above, salient words are selected from each rhetorical section because they are statistically atypical of the rest of the corpus. They 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. For comparative purposes, salient words which enjoy a higher rank in the PSC than in the Cobuild corpus are underlined. (Statistics for each section are provided later. Only five grammatical items are salient in Titles):

Table 8. Salient Grammatical Words in Rhetorical Sections of the PSC.

	Titles	**Abstracts**	**Introductions**	**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	-	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

It can be seen that some sections are more 'Cobuild-like' than others. Paradoxically, 35 of the 55 words set out in the table above are in fact relatively more frequent in the Cobuild 1987 corpus than in the PSC (as detailed in section 2.6 above). Patterns attributed to Cobuild items may represent a 'general language' quality of that rhetorical section, although as we demonstrate below, their use in fact changes significantly in the corpus. Perhaps not surprisingly however, Introduction and Discussion sections have a more 'general language' vocabulary, while the salient items in Titles and Abstracts seem to be further away from general usage. Salient words that are more frequent in the corpus (in Titles and Abstracts) presumably have phraseological patterns which move the corpus as a whole away from the general language. This sense of distance is of course a convenient metaphor: the real difference lies in the high density of use of such items as prepositions in these sections. Such features of language are noted in the analyses set out below. In summary, when grammatical items are analysed in the corpus, we are characterising a particularity of the rhetorical section that sets it apart from other sections, not necessarily one that sets the corpus apart from Cobuild or the general language. Some words, such as 'between' have a higher rank in the PSC but are relatively stable across the corpus: they are therefore not covered this kind of analysis.

In the following sections, I have set out grammatical items which are salient in several sections in alphabetical order in order to immediately compare the behaviour of an item from one section to the next (such as *is* which is salient in Introduction and Discussion sections). Secondly, certain items are very highly significant for that rhetorical section only, and can be

more usefully described in a general discussion of each section as a whole. The following tables indicate the order in which I have conducted these two analyses:

Table 9: Repeated Salient Words Sorted by Item

	Titles	Abstracts	Introduction	Methods	Results	Discussion
after				*	*	
and	*			*		
did		*			*	
for	*			*		
have			*			*
in	*	*			*	*
is			*			*
not					*	*
of	*	*	*			
that		*				*
there		*			*	
was		*		*		
we			*			*

Table 10: Unique Salient Words Sorted by Section

Title	Abstract	Introduction	Methods	Results	Discussion
on	but	been	were	no	be
	these	has	at	had	may
	who	such	then	the	our
	both	can	each	when	this
		it	from	all	
		to	with		

Each one of these items is analysed as a node word below, thus *has* and *have* are analysed separately (it is worth noting here that each word form has a sufficiently different set of collocates to justify this separation, a point defended in our discussion of the lexico-grammar, above). These salient words are analysed below with the data that motivate their selection (these figures can also be seen in the Appendices). I have attempted to limit the number of examples of collocation to five, although there is some variation in this. With long examples I have sometimes had to omit all other elements except the heads of complex nominals or omit modifying words which did

not fit into the span (for example, a long set of technical pre-modifiers placed before a significant collocate of the node word).

One specific finding which emerges from the corpus needs to be signalled here before I set out the data in full. There is a strong tendency for collocations to cluster around lexical items that share similar semantic characteristics. Four process types appear to predominate in the corpus data. They are listed here from relative proximity to the scientists (research processes) to relative distance (biochemical processes):

a) RESEARCH (cognitive, verbal processes) or 'metacomments' about research itself, and which characterise the writing activity or act of observation that the researchers are engaged in (for example, from the Medline corpus: *study, evaluation, case, comparison, analysis, detection, characterisation, assessment*).

b) CLINICAL (material, behavioural processes) include the medical or methodological processes carried out specifically by the scientists in experimentation: (e.g. *treatment, therapy, care, management, resection, injection*).

c) EMPIRICAL (relational, material, perceptual processes) refer to theoretical models or express quantitative observations and the behaviour of data (*effect, role, risk, influence, use, relevance, stability, increase*).

d) BIOCHEMICAL (material, behavioural processes) identify the technical biochemical interactions and entities observed by the researchers: (*expression, infusion, synthesis, hydrolysis, induction*).

I find below that so called 'regular' phraseological units typically restrict the semantic components of the phrase to one of these process types (or even a subtype). In other words, one of the defining characteristics of each process type is that they occur in complementary distribution to each other. This is in effect the principle behind the original Cobuild dictionary: senses are defined by collocational or even grammatical behaviour. I use this classification to describe the global characteristics of a phrase but emphasise here that these categories emerged initially from the corpus analysis and need to be considered in their phraseological environment.

It should also be noted here that I make reference to clause structure often in terms of Hallidayan grammar (1985), including terms such as relational (copular) clauses and material (transitive) clauses, adjuncts (sentence modifiers) etc. The scientific processes: biochemical, clinical, empirical or research also closely relate to Halliday's transitivity processes (material, relational, verbal, mental, behavioural...). For example, most research processes correspond semantically (if not phraseologically) with Halliday's mental or verbal processes.

2.The Phraseology of Salient Items

In this section I set out alphabetically those grammatical items which are salient in more than one research article section. Their relative rank of salience in relation to the *Wordlist* comparison is included in brackets.

2.1 AFTER$_1$ (Methods salient word 9).

We have seen above that in a general lexical comparison between the PSC and the Cobuild corpus, prepositions emerge as the most significantly frequent items in science writing, whereas auxiliaries and modal verbs, conjunctions, pronouns and determiners appear to be less prevalent. This suggests that the research article genre differs from the general language at a basic grammatical level in nominal groups (in which prepositions play a key role), phrasal / prepositional verb usage and the use of sentence adjuncts. The phraseology of 'after' is important in Methods sections in the expression of time. The preposition does not however head a time-related PP (preposition phrase), but instead introduces a clinical process performed before the action indicated by the verb. The methodological procedure is thus presented in reverse order in the sentence. Some typical examples include:

{Clinical process}	**after**	**{Clinical nominalisation}**
were added 24 hours	**after**	amputation
were killed 26-30 days	**after**	injection
cultures grown 3 hours	**after**	the start of chemotherapy
regimes administered several hours	**after**	heating at reflux
l-action was applied for 2 hours	**after**	drug administration

After tends to be introduced by passivised clinical or experimental interventions such as *obtained, added, killed* (its 3 most frequent lexical left collocates). This is markedly different to its use in the general language, where *after* more frequently introduces a time expression in narrative (according to Cobuild the most frequent uses include *after two days, after a while*: these are more frequent, but of course the preposition enters into many other patterns). Furthermore, we can see that alternative time expressions in the PSC take on a rather different phraseology. For example, if a specific

time reference is missing in the left-hand expression, *after* is usually intensified by 'immediately':

removed	immediately	**after**	sacrifice
returned to their cages	immediately	**after**	surgery
saline was removed	immediately	**after**	surgery
excised	immediately	**after**	exposure
cut into two parts	immediately	**after**	the cyclophophanine infusion

These expressions also provide numerous euphemisms for killing experimental animals (as in the example *after sacrifice*). Various euphemisms of this sort emerge in our corpus data below.

AFTER$_2$ (Results salient word 6).

In Results sections, *after* is used predominantly in the phrase <*after treatment*> (more than 50 occurrences). Apart from time periods, *observed* is the most frequent left-collocate, and in many examples *after* takes on its more usual general language function introducing time phrases:

the resistant phenotype	**observed**	**after**	10 min. dilution time
the phenotype was	**observed**	**after**	2 days cultivation
the resistance was	**observed**	**after**	4 weeks of treatment

This might be taken as a small move in the direction of general language style. The lexical phrase <*after adjustment for*> also becomes prevalent in Results sections and is used sentence-initially (in the terminology of theme-rheme analysis) in a complex topical theme. As I point out in my specific discussion of Results sections below, much of the recurrent phraseology of this section has to do with rephrasing. In this case, the expression reformulates a variable and passes over or summarises a complex set of calculations:

<**After** adjustment for>	other factors, we
<**After** adjustment for>	birth weight
<**After** adjustment for>	this additional variation
<**After** adjustment for>	tumor stage

<**After** adjustment for>	the same factors

2.2 AND_1 (Title salient word 4).

Conjunctions are perhaps the least likely candidates to display collocational properties. Yet *and* appears in a number of relatively predictable collocational frameworks throughout the corpus, for example: *combined* {research process / clinical process} ***and*** (research process / clinical process}, where the word *combined* appears to function in Titles as an additional intensifier:

combined	presentation	**and**	discussion.
combined	chemotherapy	**and**	evaluation.
combined	evaluation	**and**	comparison.
combined	diagnosis	**and**	management.
combined	modality advance radiation in children	**and**	radiotherapy.

Since *and* is a salient word in Titles, it presumably has a significant role in the presentation of data. While *and* is treated in general language as a conjunction signalling similarity or connectedness in longer stretches of discourse, in research article Titles it is primarily used to signal causality. In other words, the conjunction joins items that may be construed to be worthy of scientific enquiry and has the pattern: {disease related cause} **and** {disease}:

diet	**and**	cancer
dementia	**and**	cancer
colorectal cancer	**and**	genes
gastric cancer	**and**	metastases
the role of color Doppler US	**and**	prostrate cancer

A longer expression on the same semantic lines appears to be triggered by an empirical process item (such as *link, differs, relates, relationship*) and involves a collocational framework *between* _ ***and*** or {empirical process} {*between*} {disease related phenomenon} **and** {disease}:

gene expression differs	**between**	species	**and**	malignant tissues
link found	**between**	smoking	**and**	risk of cancer
relationship	**between**	gene amplification	**and**	long term malignancy
relationship of GerB expression	**between**		**and**	endometrical cancer
Prototatic TRH relates peptides	**between**		**and**	high cell count

It is notable that these Titles (derived from the *Medline* subcorpus) involve non-finite and finite clauses, which are as we have noted above a novel characteristic of Titles in developmental biology. Besides relating previously unrelated causes of disease, relationships are also established between scientific disciplines:

The relation	between	clinical	**and**	histological outcome
Bridging the gap	between	research	**and**	clinical practice

Similarly *and* links complementary items belonging to a limited class of related items in the collocational framework *in _ and*

(cancer)	in children	**and**	adolescents
(patterns of breast cancer)	in Asian	**and**	Caucasian women
(clinical applications)	in prognosis	**and**	disease monitoring.
(mechanism of action)	in disease	**and**.	therapy

Such a framework of complementary listed items also appears to be initiated by left-collocates of *'of'* in expressions such as *'Potential combination of'*. This includes research and empirical process items: *detection, comparison, impact, role, effect, levels*. This leads to a longer collocational framework of the form *_ of _ in _ and _* As can be seen in a number of Titles in Appendix 2, a general pattern emerges with the following phraseology: {general finding} of {focus of research: a biochemical entity} in {data sample}. For example from the PSC:

- Prolonged retention of high concentrations **of** 5-fluorouracil **in** human **and** murine tumours.

- Developmental toxicity **of** boric acid **in** mice **and** rats.
- Antitumor activity **of** the aromatase inhibitor FCE 24928 on DMBA-induced mammary tumors **in** ovariectomized rats treated with testosterone.
- Comparative immunology using intact fragments **of** ...anti-CEA antibody **in** a colonic xenograft Model.
- The influence of the schedule and the dose **of** gemcitabine on the anti-tumour efficacy **in** experimental human cancer.
- Characterization **of** p53 mutations **in** methylene chloride-induced lung tumors from B6C3F1 mice.

It appears that the phraseology of the framework *of_in_(and)* forces us to interpret each constituent in rhetorical rather than lexical terms. In other words, nouns which would normally be seen as part of a general semantic field have a specific role within the title. For example, *developmental toxicity, comparative immunology* and *characterization* are seen as research fields or research activities out of context, but in NP (nominal) Titles they can be considered as the main finding of the article. Terms such as *Characterization* and *Developmental toxicity* are claims as a function of being placed in thematic position within a complex nominal, but their associated meaning of result or finding is also reinforced by the appearance of other lexical items which are unambiguously empirically oriented in this position. They can be compared with *The influence of the schedule ... Antitumour activity* and *Prolonged retention* which are specific claims about effects or new data. In Titles, '*Influence'* and '*Antitumour'* express a biochemical claim about causality, while '*Prolonged'* makes an empirical quantitative claim. This can be further compared with expressions in which the second (grammatically subordinate) element is introduced by *in* and the nominal head reformulates an empirical claim: *Decreased resistance to N,N-dimethylated anthracyclines in multidrug-resistant Friend erythroleukemia cells*. Nominal patterns with *of* and *to* express a transitive relationship and are relatively fixed. They both operate in parallel to nominals with *in*. Patterns with *and* are less fixed, but operate within the overall phraseology and extra complexity within the nominal does not affect the overall pattern (as can be seen in the Titles in Appendix 2). Such patterns provide a consistent schema which places the findings of the research in thematic position when the Title is expressed nominally (and this pattern differs considerably from the many non-nominal Titles where the findings are placed more stereo-typically in sentence-final 'new' position as in *pS2 is an independent factor of good prognosis in primary breast cancer*). My claim is therefore that while these would perhaps be trivial patterns in terms of the general language,

grammatical frameworks correspond to highly meaningful phraseology within the context of research article Titles.

AND_2 (Methods salient word 7)

As with the items *'then'* and *'each'* which we see below, the statistical significance of *'and'* in Methods sections is due to the general tendency to sequence stages of clinical and empirical analysis. *And* is used in fixed expressions which can be seen as routine collocations, as in the following recurrent examples: *cut and stained, cut and mounted, cut and plated, cultured and plated , sected and stained with...treated and counterstained with removed and routinely stained with...developed and stained...* However, chronological sequence is not always respected in the phraseology, and clinical processes such as *collected* seem to be expressed as a redundant intensifier:

collected	**and**	counterstored
collected	**and**	mounted
collected	**and**	placed
collected	**and**	stored

Such unremarkable phraseology stands in stark contrast to the key role of *and* in the expression of causality in Titles.

2.3 DID_1 (Abstract salient word 8)

We have seen in the basic statistical count that verb forms, especially auxiliary and modal forms such as *did* and *have* are in fact somewhat less frequent in the PSC in comparison with Cobuild. The salience of *did* in Abstracts and Results is therefore significant, because we are dealing therefore with a phraseology that is very specific to these two sections. The modal verb *did* is only used in two ways in Abstracts: to introduce the negative *not*, and in elliptical expressions such as *<as did the> + NP...* Perhaps surprisingly, the presentation of negative results is a key function in Abstracts. Such findings are included partly to deflect possible criticism but also because empirical negative results are just as newsworthy in the discussion 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 empirical measurement or reporting verbs prevalent after *'but'* (*<but did not>... increase, decrease, show that*). Typical subjects of these clauses are quantitative empirical processes (*efficiency, correlation, the data, sample response*). This pattern differs slightly for *did* in Results sections, where negative findings tend to relate to empirical processes of causality rather than quantification. The reason for the difference in expression may be that Results sections tend to justify and explain negative findings (such as lack of causality, effect or evidence) while Abstracts state data-related results, leaving inferences about 'higher' empirical or research implications to the main text.

DID_2 (Results salient word 3)

I discuss the role of *'did'* in Results sections in the next section (under *not*). However, *did* is frequently used in two other important syntactic environments. The first after *but* is as an intensifier of a biochemical process or empirical finding (notice that in Abstracts expressions of this type involve the negative *not*):

but	**did**	appear to induce protein
but	**did**	demonstrate the presence of
but	**did**	cause a statistically significant increase in the elimination of
but	**did**	cause some increase in the levels of CYP2A
but	**did**	cease to gain weight

The second use is elliptical after the conjunction *than* and an empirical or biochemical process verb in a comparison of findings (such a discursive expression is also not used in Abstracts):

caused more weight loss	than it	**did**	in nontumour bearing mice
yielded more synergism	than	**did**	exposure to Cis PT
exerted sig. higher toxicity	than	**did**	danorubicin
produced much higher values	than	**did**	cells pretreated with both

treated mice generated more H2O2	than	**did**	C57BL mice

2. 4 FOR_1 (Title salient word 2)

'For' is a significant salient word in Titles and generally signals a specific research problem, usually a disease. Although rather infrequent in PSC Titles, *for* emerges as a salient word when the larger control corpus (Medline Titles) is compared with Medline Abstracts. In titles, *for* is used to postmodify complex nominals and has the phraseological pattern: {treatment related item X} **for** {disease related item Y}. This expression has two variants: empirical or clinical process items:

empirical item:	**for**	disease:
consequences, estimates	**for**	colorectal / breast
implications, risk	**for**	advanced ovarian
risk factor	**for**	... cancer

clinical item:	**for**	cancer of the liver...
diagnosis, radiotherapy, resection	**for**	
chemotherapy, screening, therapy	**for**	
surgery, uretoscopy	**for**	

In the larger Medline control corpus of titles, two thirds of expressions of this sort are placed in thematic position as in *Bioreversible protection for the phospho group:....* in a similar results-related pattern to the one described under *and. For* is thus not widely used as an adjunct in this part of the research article.

FOR_2 (Methods salient word 5)

In Titles *'for'* is used in a number of expressions to link causality and disease, whereas in Methods sections it expresses a stage of analysis within the methodology, for example:

the primers were	used **for**	amplification

the procedure was	used **for**	calculating the CI values
the probes were	used **for**	characterization of antibody
the supernatant was	used **for**	comparisons
the test was	used **for**	evaluation of patients

A particularly regular phraseology emerges in the expression *'examined for'* which is effectively a prepositional verb with the phraseology {animate donors / cells} <were examined for> {visible disease-related item}:

Five animals	<were examined **for**>	external defects
the animals	<were examined **for>**	soft tissue...abnormalities
Livers	<were examined **for**>	grossly visible lesions
donor organs	<were examined **for**>	visceral defects
Live fetuses	<were examined **for**>	gross defects
...carcasses	<were examined **for**>	malfunctions
Cell markers	<were examined **for**>	skeletal malformations
...cell lines	<were examined **for**>	malformation and variation

Such a regular phraseology demonstrates the effects of semantic prosody. For example, in the following expression, *The heads were senally sectioned and examined for RT activity*, we must assume that *'RT activity'* is evidence of a disease-related defect on the basis of the more general phraseology. It is worth noting again at this point that when such related but disparate items are observed in a regular phraseology they are seen as a collocational cluster.

Similarly, the adjectival complement expression <*eligible for*> is used to signal the relevance of certain data and collocates with study:

fifteen patients were	<eligible **for>**	entry into the present study
the control group	<eligible **for>**	the study
In order to be	<eligible **for>**	the study
two groups were	<eligible **for>**	the present study

2.5 HAVE$_1$ (Introduction salient word 3)

The significance of *have* (and *has*) in Introduction sections confirms many intuitive findings expressed in previous ESP research. In general the perfect together with extraposed expressions in *'it has been seen that'* is a conventional way of reporting present research processes, while the present tense, as we see for the item *'is'* below is paradoxically used to report 'given' or 'past' biochemical facts. Over 55% of the instances of 'have' in the corpus are involved in research reports in *'have been'* (discussed below). Of the remaining instances, the most common uses of the verb are as auxiliary in impersonal summaries of previous research as in *'has received'* / *'have received (little, much) attention'*, and also *'have attracted (much, a lot of) debate, attention'*. A particular phraseology is associated with the verb '*show'*, this time used in the active verb complement expression: <studies have shown that> {biochemical result}:

Randomised clinical studies **have** shown that EPX is equivalent to MTX
Immunological studies **have** shown that oral feeding in drink water correlates with several colonic cancers.
Some studies **have** shown that there is considerable heterogeneity
Earlier studies **have** shown that some activity mutation in ras genes are specific.
Previous studies in this laboratory **have** shown that semiempirical and ab initio methods can be coupled...

The only exception to this pattern is the replacement of 'studies' by the names of other researchers (*Bardwell and Cheng have shown that, Tanish and co-workers have shown that* etc.). A similar and important use of the verb is introduced by *'we'* (except that the prefers verb is 'found': *'we have found that*') but this change in collocational behaviour is discussed below under *'we'*. These general observations are in accordance with previous research on tense (Heslot 1982, Salager-Meyer 1992). However, *have* is not only used in the PSC in the direct reporting of past research but also in the expression of subjective judgements. The third use of the verb does not report previous research directly but expresses established facts in terms of positive or negative evaluation (the bracketed words are non-optional evaluations):

... **have**	a {profound} enabling effect
... **have**	a {good} prognosis

... **have**	a {high} glycolytic rate
... **have**	a {high} prognosis potential
... **have**	{poor} capacity
... **have**	{poor} oral availability
... **have**	{significant} role
... **have**	{totally different}molecular framework
... **have**	{well-documented} effect

It is noticeable throughout the corpus that present tense simple relational clauses of this type (involving *has, have, is, are*) almost always involve subjective or evaluative expressions. Simple expressions of relation without some explicit evaluation are rare. This is markedly different to patterns of usage in the general language. The *Cobuild* dictionary does not lists evaluation as a main use of *is*. It appears therefore that simple relational uses of *have* often tend to be possessive, while *is* is often used in more impersonal grammatical constructs, such as extraposed projections (*it is safe to*).

HAVE$_2$ (Discussion salient word 10)

In Introductions *'has, have'* are most often used with specific expressions of past research reporting *'have led to debate / has attracted attention'*. In Discussions, more specific research processes are more emphasised. Although most research is expressed actively in terms of *we* (see '*we'* below), passivised reports of research processes are the next most frequent use:

have been	detected
have been	found to be
have been	identified in
have been	reported to
have been	shown to

Another less dominant pattern involves reports of previous research similar to that expressed in Introductions (the pattern *have _ that* can be seen to form a consistent collocational framework with mental or verbal expressions of research):

previous studies	**have**	shown that
we	**have**	reported that
we	**have**	found that
clinical studies	**have**	demonstrated that
experiments	**have**	suggested that

And as in Introductions, attributive relational processes expressed by 'have' are used frequently to express evaluation, although this time in relation to quantitative or specific results reported in the research article rather than prior facts:

Biochemical report		**Evaluation**	**Bio - / Empirical process**
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 abilities
ragments have been reported to	**have**	superior	localisation abilities

2.6 IN_1 (Title salient word 5)

'In' is salient in four rhetorical sections in the corpus and presents us with the opportunity to test whether phraseology is consistent throughout the corpus. As noted above, prepositions appear to account for many of the major differences in vocabulary and style between the PSC and the general language (at least in terms of a comparison with Cobuild). The highly frequent prepositions *in* and *of* in the corpus are thus key to an understanding the fundamental phraseology of the genre. In Titles *in* functions as a prepositional phrase functioning as either modifier or complement in complex nominals (we have seen one use under *and* above). There are two distinct semantic patterns:

1) In modifier expressions, the left collocate is a biochemical process and the right collocate a clinical or biochemical entity. Where the head of the left phrase is not the immediate collocate, the head item is usually an empirical or clinical process. It is noticeable that for each left-collocate, a more or less

limited pattern emerges to the left again of this item (for example, gene **expression**). Head items are noted in italics:

Biochemical process			**Clinical or biochemical entity**
changes in distribution of	*cancer*	**in**	human, liver [etc]
intake and risk of	*cancer*	**in**	children, primary care
improved detection of breast	*cancer*	**in**	group practice, women
determination of screening for	*cancer*	**in**	rats, Singapore,
surgical therapy of prostate	*cancer*	**in**	the elderly, aged patients
gene	*expression*	**in**	scrotal contents
receptor gene	*expression*	**in**	breast CYP1A1
growth	*factors*	**in**	Cancer
prognostic	*factors*	**in**	colorectal cancer
Expression of trypsin and other	*factors*	**in**	gastric carcinoma
p53-like..,	*factors*	**in**	HB carcinoma
p53 expression and other	*factors*	**in**	breast cancer
diethyl analogue	*cell lines*	**in**	Culture
growth-regulatory	*cell lines*	**in**	a p53 pathway
human bladder cancer	*cell lines*	**in**	Protein
larger auxiliary	*metastases*	**in**	obese women
colorectal adrenal	*metastases*	**in**	patients with (cancer)
breast cancer	*metastases*	**in**	meginoma
evaluation of...hepatic	*metastases*	**in**	patients
prediction of auxiliary lymph node	*metastases*	**in**	tumour-bearing animals

The only exception to this pattern involves the modifier (of X) *in patients with*:

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

2) In complement expressions, the left collocate is an empirical item for which a statistical significance or medical potential is signaled in the Title. While the first pattern for *'in'* suggests a general tendency for the qualifying phrase to specify the disease (or the subjects in which the disease is to be found - a 'spatial' metaphor common in the general language), the right-collocate in the second pattern completes the semantics of the left-collocate. Right collocates are not clinical samples, as in (1) above, but empirical data sets:

Empirical item		**Empirical data set**
Significant	change **in**	levels of specific in vitro residue
significant	changes **in**	cytokyne levels
highly significant	change **in**	levels of stromal antigens
	change **in**	cachexia mortality
	change **in**	distribution of histogenic type
potential	role **in**	human disease
possible	role **in**	the metastatic process
suggests a	role **in**	tumor production

The basic distinction between *in* 1) and *in* 2) echoes Sinclair's observations of *of* in the general language. In the first case, the phrase after *in* functions as semantic support, whereas in complement expressions the prepositional phrase is the semantic focus of the entire phrase (Sinclair 1991:82-83).

IN_2 (Abstract salient word 5)

The spatial metaphor of *in* in Titles is not prevalent in the rest of the article. *'In'* in Abstracts is used in three semantic patterns (the most frequent first).

1) as nominal modifier in expressions of measurement (*significant increase in toxicity, reduction in levels, differences in cytotoxicity, decrease in uptake*)
2) as verbal modifier in attributive or relational clauses of biochemical process (*accumulates in, is low in, resistance was narrower in the cell*) and as a phrasal element in research processes (*observed in, detected in*) or empirical processes (*role in, resulted in, used in*).
3) in an adjunct, introducing research with *this* (*in this study/ trial/ phase 1 study/ report...*).

In Abstracts, *in* also introduces non-finite relative clauses where given information on a chemical process is bundled in with the original information such as *introduced in, involved in, implied in* (as in: *this is a novel approach to adaptive resistance involved in the expression of ras oncogene*). In Titles, it can be seen that the majority of uses of *in* are determined by the right collocate (*in* therefore completes the meaning of these expressions while functioning as a 'spatial' modifier of the left-hand expression). In Abstracts the spatial use of *in* is largely supplanted by a less specific meaning of the prepositional phrase (a general biochemical / empirical process) and is determined by the left-hand collocate. This also corresponds with the use of the determiner *the* (largely absent in the right-hand collocates of *in* in Titles) as in : *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 Abstracts 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*). This pattern appears to revert to the use of *in* in Titles.

It is likely that reference and other discoursal factors have a role to play in this distinction although 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 idiomatic in

certain specific semantic domains in the general language, then it may be that determiners are also constrained by prepositions in the ESP.

IN_3 (Results salient word 2).

In is used in three types of phrase in Results. The first is to indicate positive results which usually involve a higher experimental score or increased amount of measurement. This can be contrasted with the negative results which usually lack 'direction' (higher or lower score), and usually indicate only the relevance of the result to the empirical model ('directionless' findings tend to be reported in Abstracts, as seen below). The second pattern is closer to the spatial metaphor of *in* in Titles, indicating where a specific biochemical process was found / observed in the bodies of patients or subjects. A third pattern takes the form of a research process verb + preposition functioning as a cross reference to another section of the article. The first and the third patterns are specific to Results sections.

In the first pattern, the most typical use of *'in'* is to express data direction (*increase in, increases in*: 61 occurrences) after either a semi-technical empirical verb such as *'yields, expressed, produced'*: {empirical process} a/an {specific data shape} increase **in** {measurable, often disease-related empirical item}:

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

One phraseology in particular becomes prevalent in Results sections in which the verb *yield* is consistently followed by a post-nominal quantifier: *<increase in the level of>*

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

Another frequent expression in the first pattern involves the empirical process *'resulted in'* in which the direction of the data is emphasised by some intensifier: {clinical process} resulted **in** {intensifier} {empirical measure / biochemical process}. Unlike the *yielded* phraseology, this expression generally allows for very explicit modality (if no explicit evaluation is expressed, then a determiner or similar expression to the first pattern is used):

Biochemical process		**Evaluation**	
analysis	**resulted in**	marked	increases
protocols	**resulted in**	significant	deaths
concentrations of dry MM	**resulted in**	negative	induction
The same dose of DXR	**resulted in**	strong	synergism
Since increasing the dietary BORA	**resulted in**	total	loss of oral viability...

The writer may also choose to express positive results as a relation (*is, be, were*) with *higher*. Such a phraseology is oriented towards an evaluation of change in biochemical data (in animals or cells): {empirical measure} is {empirical evaluation} higher **in** {animate material}:

tended to be		higher **in**	dogs treated with 30mg
peak level is	markedly	higher **in**	tumor cell lines
drug level is	consistently	higher **in**	animals
leucocyte count is	significantly	higher **in**	the liposomal DXR groups

5FU concentrations were2 times		higher **in**	animals necropsied at

This is related to the second, spatial use of 'in' in Results sections, in which the preposition introduces a biochemical. In some cases, as in the last examples, the biochemical entity is a data set itself. For example, *'in'* is used in the basic comparison of results where the data sets are expressed as subjects or patients:

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

A more typical spatial metaphor pattern involves technical biochemical processes including the expression *'in vivo'* (although this is a Latin expression, its grammatical profile is similar to other modifiers or adjuncts introduced by *in*). Various collocational expressions emerge in terms of the spatial metaphor. *'Activity'* for example usually takes place *in organs*:

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

'Concentrations' are only found however in *'tissues'* or *'tumours'/ 'tumors'*:

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

		from the patient

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

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

Mutations in turn are typically detected ***in*** *genes* (*the p53 gene, exon 6 of p53, k-ras exons, H-ras gene*). An alternative wording is to premodify the mutation with a gene classifier, thus enabling it to be detected in *tumours* [variation in spelling here indicates the use of British spelling in such journals as BMJ, BJ, etc.]:

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

The spatial use of *'in'* also reveals terminological consistency within right-hand collocates. For example, only nude mice are used for skin grafts:

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

while frameworks with other common lexical items also reveal the terminological properties of related words. For example, *tumours* are associated with a variety of physiological locations (from *genes* and *cells* to

organs) as well as a range of conditions (*benign, necrotic, malignant*), while *cancers* are named in terms of larger organs and are less frequently mentioned. *Carcinomas* are generally limited to the expression of cellular cancers:

In	benign, breast, clear-cell, colon, colorectal epithelial, invasive, malignant, murine, necrotic, p53-negative, primary, renal cell Ta-Ti, Various	**tumour/s...**
In	Bladde, breast colonic, colorectal lung, oesophageal, pancreatic...	**cancer**
In	Basal-cell, Cervical, colorectal, hepatocellular, human cell, invasive, squamous cell	**carcinoma/s...**

Interestingly, while the Latin *'in vivo'* is often used as a sentence adjunct, its complementary expression *'in vitro'* tends to be used as a premodifier in noun groups, and so we get the following expressions (in such usage *in vitro* functions as a single lexical item - as such *in vitro* is not as clear-cut a case of *in* as *in vivo*):

The	**<in vitro>**	antitumour activity
The	**<in vitro>**	culture
useful	**<in vitro>**	growth
various doses of	**<in vitro>**	results
PKC activity of the	**<in vitro>**	system

The third overall use of *in* is a text-referencing pattern, typical of Results sections. This usage accounts for the most frequent lexical left-collocate of *in*: *'shown in'* (34 occurrences). The use of the present rather than past passive is noticeable in the following examples:

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

A range of similar research-writing verbs play a similar role:

clinical details are	detailed **in**	table X
samples are	given **in**	fig. X
doses given are	illustrated **in**	table X
grain counts are	listed **in**	fig. X
these results are	plotted **in**	table X
values are	presented **in**	table X
NMR plotting is	summarized **in**	fig. X

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

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

The expression *'as seen **in'*** is also involved in a longer fixed expression observed in two structural chemistry texts:

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

Finally, the use of *'in'* in lexical phrases in Results is more varied than for the other prepositions we observe in the corpus, and we note here briefly such expressions as *in addition, in all, in comparison, in contrast*. This suggests

that there is more explicit signalling in Results sections, although this is somewhat terser than the kinds of expression encountered in Discussion sections.

IN$_4$ (Discussion salient word 6)

To summarise the uses of *'in'* so far: in Titles, expressions after 'in' modify some biochemical item or process (*metastases in, expression in, growth in*) or complement an empirical item (*role of... in, change in*). Such patterning constitutes important evidence for grammatical and semantic correspondence, in other words a lexico-grammatical system. In Abstracts, we noted mostly nominal reformulations of quantitative results and a number of expressions involving empirical quantification (*increase in, decrease in, reduction in, difference in*). In Results sections the use of *in* extends to more complex forms of quantification, a spatial use with biochemical entities and the use of lexical phrases and cross references to other parts of the research article. In Discussion sections the tendency is again to express empirical shapes and directions of data (the most frequent pattern) and causal relations (the second pattern). A third pattern involves research processes, and a fourth comprises large numbers of discourse markers. Such increasing variation in the phraseology of a single grammatical item supports a general observation that the final sections of the research article become increasingly stylistically diverse.

The role of the Discussion section also returns to explanation, in a similar mode to that of Introduction sections. Thus the fixed expression <*play a role in*> becomes a significant phrase in Discussions where some degree of explicit evaluation is often present:

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

Biochemical items are described as (spatially) *'present'* and stated as implicitly observed facts:

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

A similar pattern is seen in verb- or adjective- complement expressions *is reflected in, is similar in*, and *is visible in*. Unlike many simple present tense use of relational verbs in the corpus, adjectives used in complement constructions are rarely accompanied by explicit evaluation. This represents a general move away from quantified observation in Results sections to qualified empirical observation. Specific results are reformulated or identified as *'found'* or *'observed'* in the passive (*similar response was observed in this study, LOH has already been found in all renal tumours*). Finally, *'in'* tends to be used in complex NP-complement prepositions. These take the form of collocational frameworks where the whole expression functions as a discourse marker. For example, *'in _ to'* allows for contrasts:

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

while *'in_ with'* signals that findings have or have not been replicated elsewhere:

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

2.7 IS_1 (Introduction salient word 4)

The verb *is* is fundamental to the phraseology of Introduction sections. As with the relational verbs *'has / have', is* is used to signal explicit evaluation. In the PSC, the phraseological patterns of *is* are (in order of frequency):

1) Introducing an extraposed adjectival complement clause: *It is* {modal item} *that* {treatment related item X} {biochemical / empirical process}:

It **is**	unlikely that	(X)	does not express its gene products
It **is**	possible		plays a key role
It **is**	assumed		increases in direct relation to
It **is**	possible		needs to be well separated
It **is**	conceived		differs at the level of tumor production
It **is**	well known		can be modulated
It **is**	relevant		is the main source of circulatory...

2) Introducing an extraposed adjectival non-finite complement clause (limited to three adjectives) *It is* {modality item} *to* {research process}

It **is**	possible to	identify TAAs that allow
It **is**	necessary to	assess the cell differentiation at this stage
It **is**	important to	obtain structural information
It **is**	possible to	construct a series of... structures
It **is**	necessary to	identify mechanisms of drug resistance
It **is**	possible to	repeat measurements
It **is**	necessary to	establish whether
It **is**	important to	study forms of the enzyme

3) Introducing an adjectival or verbal non-finite complement clause: a {Biochemical process} *is* {research utterance} *to* {biochemical process}. There are only three possibilities for this type of expression. These are alternative expressions, indicating decreasing levels of certainty through modulation in verb group complexes (a type of grammatical metaphor):

Hyperphasia	**is**	**known to**	inhibit
Enzymatic...	**is**	**known to**	processed generally via

HPV 16 E6	**is**	**known to**	bind p53
metabolism inc-cells	**is**	**known to**	be proton-elevated
{Biochemical}	**is**	**likely to**	be involved in...
	is	**likely to**	arise from differences in...
	is	**likely to**	differentiate in many cells
	is	**likely to**	attract factors from hepatocytes
{Biochemical}	**is**	**thought to**	be a major factor in
	is	**thought to**	determine cell cycle
	is	**thought to**	act via ....crosslinking
	is	**thought to**	be one of the most important

4) A fourth use involves equative relational clauses: where X is a specific {biochemical process or item} : *Pancreitis, resistance to therapy, BORA, the Winsford deposit...*

X	**is**	a/an common	predictor	X	**is**	a/an important	target
X	**is**	an appealing	alternative method	X	**is**	an effective	inhibitor
X	**is**	a critical	parameter	X	**is**	a potent	derivative
X	**is**	a major	Sign	X	**is**	a potential	agent
X	**is**	an imperfect	route	X	**is**	a strong	inhibitor

Impersonal existential clauses are also used to express explicit evaluation:

there **is**	a	strong motivation
there **is**	a	substantial difference
there **is**	a	positive correlation
there **is**	a	clear need
there **is**	a	significant possibility

When *'is'* is used in equative relational clauses (i.e. where the verb simply identifies one token as another), the element of evaluation is transferred to a notion of 'measure' or 'causality' as in the fixed expressions *'is one of the*

most...is one of the main causes of'. In attributive clauses, on the other hand, disease- and treatment- related items have stereotypical patterns Only disease related items, for example can be 'associated with':

toxicity	**is** associated with
weight loss	**is** associated with
aberrant cell proliferation	**is** associated with
an exogenous retrovirus that	**is** associated with
overexpression of p185 gene	**is** associated with

Conversely, only treatment-related items are expressed in comparison, using 'more' {+ empirical property}:

target orientation	**is** more	efficient
MTX as an inhibitor	**is** more	efficacious
a new foliative agent	**is** more	localised
this choice of prodrug	**is** more	popular
antitumour activity	**is** more	table

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. So phraseological patterns correlate according to some convention with the common semantic categories naturally involved in the research. This is complicated however by the varying phraseologies of different word forms. I note later that these patterns do not correspond with the use of *'was'* (in Methods and Results sections).

Is also reveals a limited set of items which can introduce nominal complement (projecting) clauses (known as 'fact clauses', as in *the fact is that*: Halliday 1985:244). Fact clauses in the corpus are almost always empirical and premodified by some degree of evaluation. The following list gives all the possibilities:

A 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...
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

However, there is one important exception to the evaluative pattern for *'is'*. In the Introduction corpus, when the researchers are saying that something *is not* something else, explicit evaluation becomes more implicit:

Although its sensitivity to ATP	**is** not yet proven, mouse stamen have been examined...
Although cholesterol	**is** not fully responsible for the formation of liposomes, it is often used in pharmaceutical liposome formulation
Although the regulation of MyoD1	**is** not fully understood [it and others] appear to perform critical functions
Despite massive lipid mobilisation, the plasma level of these metabolites	**is** not elevated in the cachectic state...
While p52 expression	**is** not detected, it is unlikely that overexpression is related to LMF factors outside the cell.

Again, the negative relates to empirical or research processes in similar expressions to the pattern *'Although it has not been shown that'* described under *'been'* below. To summarise, affirmative phrases with *'is'* almost exclusively express modality in terms of empirical processes. Negative expressions of relation, however, deal with the full range of research, empirical and biochemical processes. In both patterns, the distinction between various genre-specific process types {biochemical, empirical, research} appears to coincide (in some cases) exactly with syntactic patterns.

IS_2 (Discussion salient word 4).

'Is' is a salient word in Introduction and Discussion sections. In Introductions, the major patterns were seen to be:

1) It **is** {empirical item} that {biochemical process}

2) It **is** {evaluated empirical process} to {research process}
3) {Biochemical process} **is** {research utterance} to {biochemical process}

In Discussion sections, as with other grammatical items the patterns are more distributed across a range of expressions, have a greater emphasis on research processes and evaluation and have in some cases different lexical components:

1) It **is** {evaluated empirical item} that {biochemical process}
2) It **is** {evaluated empirical item} to {research process}
3) There **is** a {evaluated empirical item}
4) (This) **is** {attributive research / evaluative process}
5) {Research process} **is** not {evaluative}
6) {Biochemical process} **is** {biochemical / empirical process}

Projecting (verb / adjective complement) clauses are still prevalent in Discussions however the range of adjectives and participles involved is somewhat more restricted. Whereas most projection in Introductions is related to modality and hedging, projections in Discussions sections emphasise more affirmative evaluation:

It **is**	interesting that	
It **is**	apparent that	
It **is**	clear that	
It **is**	most likely that	

Less affirmative modality is expressed by extraposed non-finite (*to*) clauses (*'It is* AP *to'*):

It **is**	possible to	screen for cell lines
It **is**	difficult to	determine influence
It **is**	important to	mechanistically link
It **is**	unlikely to	be the case that

A fixed lexical phrase is used to introduce a new research gap: *<little is known about>* and this differs from the use of 'known' in Introductions (*X is known to*):

Little **is known about**	hepatic regulation
Little **is known about**	hepatocarcinogenesis
Little **is known about**	the way the relationship helps changes in immune tests
Little **is known about**	the physiological importance of ... endothelin
Little **is known about**	the behaviours of p53 gene

In Introductions, negative relational processes are concerned with negating the empirical relevance of biochemical processes (*sensitivity is not detected, cholesterol is not applicable*). Here the tendency is to express a negative evaluation of the research process:

It	**is not**	yet clear (x5)
The latter finding	**is not**	convincingly determined
the present study	**is not**	feasible
The reason for this unexpected result	**is not**	known
Sampling required for analysis	**is not**	very defined
The functional implication	**is not**	surprising
This strategy	**is not**	very different

When results are expressed after expressions of biochemical processes, some degree of quantification is expressed as an adjunct: {biochemical entity} **is** {biochemical process: expressed} {quantification}:

the polypeptide	**is expressed**	at a very low stage of differentiation
activity	**is expressed**	only in a minority of the tumor cells
peripherin	**is expressed**	at high levels
protein	**is expressed**	as micromoles
tumor size	**is expressed**	by diameter

There are also a number of explanatory expressions where a biochemical process of disease or treatment is empirically related to observed data:

hypoglycaemia	**is** associated with	considerable increase in
The tumor mechanism	**is** associated with	acquisition of t-cell properties
The MAC tumor	**is** associated with	increased lactation
MOR phenotype	**is** associated with	enhanced stability
Oncogene p185	**is** associated with	internalization of bleeding

damage	**is** due to	observed alterations
induction in the liver	**is** due to	direct action
The presence of normal bones	**is** due to	direct interaction
Suppression	**is** due to	subsequent incubation
The positive reaction	**is** due to	the effect of.. filters

However, these patterns contrast with *'is related to'* which has as subject an empirical observation which is related to more specifically biochemically oriented items. Unlike empirical expressions in Abstracts and Results sections, and as noted above in the phraseology of *in*, these phrases deal more with qualitative explanation than with quantitative measurement. The following pattern is shared by less frequent expressions (*'is present in', and 'is responsible for'*):

risk	**is related to**	ethnicity
efficiency	**is related to**	stabilisation
the cause of toxicity	**is related to**	spasmodic polypeptides
presence of protein	**is related to**	expression of class III antigens
frequency in some tumor samples	**is related to**	the schedule of administration

2.8 NOT_1 (Results salient word 4)

As might be expected, the phraseology of *not* has less to do with propositional negation and more to do with a broader rhetorical distinction between empirical tendencies and findings (the affirmative) and empirical explanation (the negative). Examining the patterns of verbs used with *not*, we

can see that while verbs like *'show'* are used in affirmative statements to describe *'increases in'* the data, or changes of the data shape (as described under *'in'* above) negative expressions with *'show'* are used mostly to explain the relevance of data or the idea that a specific biochemical phenomenon did not take place. The implication is that in Results sections, the researchers are making a statement about causality in relation to their 'failed' or negative hypotheses but use positive statements for reporting changes in the data shape. This is contrary to the pattern in Abstracts, where negative polarity is reserved for quantitative statements (usually related to adversative expressions signalled by *but*).

The most frequent right-collocate of *not* is 'show': {biochemical entity, usually living cells} did not show {biochemical process, usually treatment related}:

controls	did **not** show	RT activity
females	did **not** show	any antitumor effect
MCR lines	did **not** show	cross-resistance
chemo-treated mice	did **not** show	greater response
the population	did **not** show	allelic loss

Similarly, the very frequent right-collocate, 'differ' emerges in a very fixed expression of findings: {biochemical process} did not differ {empirical evaluation of measurement or sometimes biochemical process} from that / those {research process}:

concentrations	did **not**	differ
bile content	did **not**	differ morphologically from that of
the consumption rate	did **not**	differ significantly from those measured
extravasation	did **not**	differ significantly from those observed
the lipolytic factor	did **not**	differ significantly from that seen in

Empirical measurement items such as: *incidence, concentrations, increasing serum levels, body weight, leucocyte counts* are all used in a similar way in a relational clause: *were **not** statistically significant*. This can be contrasted with affirmative relational clauses and uses of the verb 'show' in which researchers tend to write that data are *'increased'* or *'elevated'*.

Clearer examples of the negative in biochemical processes involve the expressions of the very frequent verbs *'express'* and *'induce'*, and this again reveals common subject-verb preferences. Cells or cell lines 'express' biochemical compounds,

the majority of cells	did **not** express	peripherin (x3 instances)
cells in this clone	did **not** express	RA activity
some cell lines	did **not** express	myocenin
only one clone	did **not** express	t-PA
the g14 cell line	did **not** express	capsid antigen

while drug therapies tend to 'induce' biochemical effects:

chemotherapy	did **not** induce	a depressor gene
lower doses	did **not** induce	any antitumor effect
CYPZA	did **not** induce	loss of weight
peptide	did **not** induce	any cytotoxicity
stronger treatment	did **not** induce	weight loss

Such biochemical process verbs have very much the same distribution as nominalisations (c.f. *induction of tumor necrosis factor*). But there are also cases in which biochemical processes are explained rather than simply observed, in which case the writers use less technical verbs such as *'cause'* and *'affect'*. For example, *'affect'* is very specifically limited to the chemical process of (cell) binding:

pre-incubation	did **not** affect	cell growth
IL 2 secretion	did **not** affect	anchorage
Those inhibitors	did **not** affect	binding
Antibiotic concentrations	did **not** affect	subsequent binding
magnetic field exposure	did **not** affect	binding capacity

In the passive the affecting medium (expressed a left-collocate above) is reformulated as a 'treatment':

accumulation was	**not** affected by	the treatment
relaxations were	**not** affected by	nitro-L-arginine at any dose
reaction kinetics were	**not** affected by	incorporation of cholesterol
excretion vomiting was	**not** affected by	the presence of ...danorubicin
weight gain was	**not** affected by	treatment with... antibodies

'Cause' is not passivised, but similarly presents a biochemical relationship albeit of a less restricted variety:

did **not** cause	mutations in the p53 gene
did **not** cause	further inhibition
did **not** cause	lysis
did **not** cause	any mortality
did **not** cause	tumorigenesis

Such expressions can be partly seen as brief claims or explanations, but can equally be seen as fixed delexical phrases (such as *take a bath, make one's fortune*). Apart from biochemical or semi-technical explanations, the negative in the Results section is also used to signal what the researchers didn't find. With *'was / were'*, we see below that the passive in Methods sections tends to be used with technical biochemical process verbs. In Results, the passive reverts to research process verbs and, at least in negative voice, is usually modal: {biochemical process} could not be {research process}:

lipophilicity	could **not** be	detected
degenerated mitochondria	could **not** be	explained
chimeric mRNA	could **not** be	related
Overexpression of p53	could **not** be	observed.

Other verbs involved in this expression are *distinguished, established, maintained*.

NOT_2 (Discussion salient word 7).

Whereas in the Results subcorpus, negative statements concerned causal relationships (*affect, cause, express*) and the general shape of the data (*did not increase*, *is not different* etc.) the Discussion sections express negative research observations. Again, unlike Abstracts, data directions are not emphasised in Discussion sections, and the emphasis is more on reformulating results than on explaining negative results. One research pattern emerges as a very regular collocational framework: 'did not {research process} any {empirical item}' and it serves to report negative results:

we did	**not**	detect	any changes
we could	**not**	find	relations
we did	**not**	observe	tumor development
we could	**not**	obtain	evidence of precursor
Early reports did	**not**	suggest	major difference

The negative also plays a key role in signalling gaps in existing research. The expression, *not known* is part of the 'end-game' of the Discussion section which allows for further applied research:

The specific source of serum To is	**not** known
The exact mechanisms of the antitumour effect of IFN are	**not** known
The functional implication... is	**not** known
Whether this is also reflected in demethylyation... is	**not** known
The nature of the inhibitory factor is	**not** known

Another important signal for future research possibilities is 'not clear' where negative findings are reformulated by higher empirical or research processes (in italics):

The reason for this *difference*	is **not** clear.	

The reason for this *latter finding*	is **not** clear.	
However, it	is **not** clear	what *differences* if any exist.
The *relationship* between gene p53 mutations and p-expression	is **not** clear.	

with one longer reformulation:

It is therefore **not** clear why cells are not able to [use] serum plasmogen.

2.9 OF_1 (Title salient word 1)

'Of' eclipses *'the'* in an Astec comparison with the Cobuild corpus, and is a salient word in Titles, Abstract and Introduction sections, thus marking its phraseology as particularly typical of technical science writing. While the use of *of* described below is somewhat complex, it is worth noting that the four or five major uses of the preposition in the PSC can be contrasted with a very broad set of uses in the general language: Cobuild, for example, lists 19 non-idiomatic uses for *'of'*.

In Titles, as in the rest of the corpus, *'of'* is fundamental to the construction of complex nominals, in particular expressions of empirical relations and quantification as well as compound nominal terminology. In Titles there are no examples of quantification (*a number of*), or support (*a group o*f). Instead, *'of's* left-collocates are nominalisations of research or empirical processes (*effect/s of* x30, *treatment of* x24, *study of* x16, *evaluation of* x15) while its right-collocates are nouns synonymous with the illness or the patient (*cancer* x69, *human* x26, *breast* x25, *patients* x18, *tumor* x15, *prostate* x13). The majority of the left-collocates of *'of'* can be divided into four groups of patterns. Research processes are the most frequent left-collocates of *of* in Titles, and typical expressions from the Medline control corpus include nominal research process titles premodified by a topic-specific specifier and post-modified by illness-related items most often involving cancer patients. The expression *'study of'* is typical:

Therapeutic	study **of**	metastasis in women aged over 40
Basic	study **of**	post-operative surgery
Comparative	study **of**	NCC-ST-439 in breast cancer.
Collaborative	study **of**	subjects participating in...trials
Case - control	study **of**	HIV-infected carriers
Immunohistochemical	study **of**	women with early breast cancer.

The research process expression -*evaluation of*- (x15 in Medline) is different in that it is seldom premodified (and is thus usually the first word of the Title), and appears to have a more limited set of postmodifiers, such as semi-technical empirical process items which are less concrete than those for 'study of':

Evaluation **of**	effects of radical resection on liver metastasis
Evaluation **of**	factors aggravating postoperative recovery
Evaluation **of**	factors affecting success of chemotherapy
Evaluation **of**	factors affecting laboratory data
Evaluation **of**	quality of life in postchemotherapy

We have seen in a number of instances a small change of expression is associated with a change in the semantic composition of the phrase. To demonstrate this we can see that the expression *'study on'* has a different phraseological pattern from *'study of'*. Left collocates are more limited for 'study on' but are more specific in terms of research activity (*case control* x5, *clinical* x3, *basic* x3, *clinicopathological* x2, *collaborative, immunohistochemical, population-based, randomized, retrospective, screening*). Right hand collocates of -*study on*- are empirical processes or items, rather than disease-related items introduced by *'study of'*:

A {research process} study **on**	clinical prediction
A {research process} study **on**	effects of continued...infusion
A {research process} study **on**	effectiveness of UFT against cancer
A {research process} study **on**	the inhibition effect of granisteron on...

A {research process} study **on**	usefulness of bleomycin in comparison with...

My claim is that the most stable elements of a phraseological opposition are important signals of the larger phraseology i.e. '{research process X} *study* ***of*** {disease Y}' on the one hand and 'A {specific research process X} *study* ***on*** {empirical process Y}' on the other. This can be seen to be an entirely conventional distinction, with little relation to any intrinsic meaning of the prepositions concerned. The distinction cannot be put down to lexical selection (or 'lexical projection' as in universal grammar), since both expressions share the same left-hand collocates. If there were some base meaning for *'of'* (as claimed by Quirk 1995) then 'Evaluation of' would not have a different pattern to other *'of'* phrases introduced by research process items, nor share a similar phraseology to 'study on'.

Clinical process phrases such as 'treatment of' and 'management of' share a similar phraseology to 'study of':

surgical	treatment **of**	solid carcinomas
combined	treatment **of**	human breast cancer
recombinant	treatment **of**	gastric cancers in Singapore
surgical	treatment **of**	breast cancer patients treated with EORTC

Of empirical processes, the phrase 'effect/s of' is the most frequent in the subcorpus and has the following phraseology: {treatment-related item X} effect/s of {treatment X} on {illness-related item Y}:

	effect/s of	chemotherapy	**on**	metastases
biphasic	**effect/s of**	aspirin	**on**	colorectal cancer
inhibition	**effect/s of**	surgical intervention	**on**	pancreatic cancer
prognostic	**effect/s of**	optimism	**on**	cancer related stress
therapeutic	**effect/s of**	somostratin	**on**	the growth of... cancer

This kind of pattern is a collocational framework can be seen to be similar in semantics to *'study on'* which in turn sometimes introduces *effects of*. A chain of phrases may be inevitable in such a conventional context, and we

find that there are many such 'collocational cascades' in the corpus. What is interesting about them is that phrases such as *'effects of'* appear to be implicit in the longer chains, or are reformulated.

An idiomatic use of the phrase 'a case of' emerges. While the word 'case' on its own is involved in the longer phraseology *'a case control study in (Brazil / Greece / Sweden)* ***of*** *(subjects participating in the Nottingham study / the blood screening programme)*', it also acts as head for 12 titles introducing specific disease-related items which are then postmodified by a response to the disease {treatment} or (in a minority of examples) an explanation of its cause:

A case of	complete response by intra-arterial injection
A case of	advanced oesophageal carcinoma treated by...
A case of	lung cancer responding significantly to...
A case of	pulmonary carcinoma which responded to treatment with
A case of	drug induced pneumonitis caused by oral etoposide.

OF_2 (Abstract salient word 3)

In the control corpus of Titles (as seen above), *of* plays a key role in nominal groups with a typical treatment-*of*-disease pattern. Such a symmetrical solution-problem pattern is expanded in Abstracts, the major difference being that while items in the title corpus tend to predict *of* with no strong right-collocates, in Abstracts there are just as many significant right-collocates, such as *human, these, was*. Another difference from Titles is that Abstracts involve the quantification or description of disease, where *of* introduces 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 either empirical or biochemical items that explain the potential treatment of the disease (*effect, role, mechanism, treatment / inhibition, synthesis... of.. drug X, doxorubicin, compounds, [disease Y]*). 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 as 'focus' rather than support (Sinclair 1991:82-83).

The 'treatment-of-disease' pattern can be seen as an overriding pattern, but within this there is considerable phraseological change. There are four different problem-solution patterns of complex stereotypical phraseology involving *of* in the Abstract: (*effect, loss, number, presence*). 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 a specific 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 qualifier, (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 more 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 whereby 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 form a separate pattern. 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 expressions of measurement-disease mentioned above, the reporting verb precedes the expression (*shows / confirms / indicates ...the presence of*, *incidence of, absence of*). The final, 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 place the study in the position of sentence theme (that is: as subject or in front of the subject in English). These patterns also suggest that choice of expression in Titles is constrained to the extent that the writer must either use measurement-disease phrases as a statement of research topic, or alternatively thematicise the results and use an expression with items such as *effects*.

OF_3 (Introduction salient word 9).

'Of' in the Introduction serves to qualify empirical process nouns and to form fixed biochemical or clinical terminology. This is the same function as in Titles and Abstracts, the difference being that the fixed expressions and collocations in the Introduction are expanded to longer stretches of phraseology. In examining the very complex phraseology of *of* in this less constrained environment, the assumption is that collocation operates at longer boundaries than the phrase. The following left / right collocates demonstrate the variety of collocation:

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

A number of longer phrases become prevalent in the Introduction and a number of phrases identified in the Title or Abstract take on a different environment. In particular we find a strikingly long collocational framework in the form of a projecting fact-clause: <the aim / purpose **of**> (this study) <was to> {+ research process} {measurable biochemical activity} (16 occurrences) :

The	aim	**of**	this study	**was to**	compare
The	aims	**of**	the present study	**were to**	examine (x3)
The	purpose	**of**	the current report	**was to**	investigate
The	aims	**of**	this work	**were to**	relate
The	aim	**of**	this series of studies	**was to**	measure uptake
The	aims	**of**	this study	**were to**	test
The	aim	**of**	the present study	**was to**	expand data
The	aim	**of**	the current report	**was to**	identify

The	aims	**of**	this work	**were to**	determine

The (missing) complements of the research processes above are measurable activities: *activation, uptake, circulatory responses, pharmokinetics in the liver, concentration of pituitary humours, p52 on mRNA expression, a possible prognostic of tumour regression.....* While in the abstract expressions involving *effects of* are generally followed by some degree of evaluation or an empirical process (*the effects **of** treatment X are demonstrated*) here the phrase occurs as complement of some research process:

{research process} *{treatment related item X}* effect **of** *{treatment X}*

assess	the adverse antitumour	**effects of**	BORA
investigate	the chemopreventative	**effect of**	boron on mice
show	the inhibitory	**effect of**	cholesterol
report	protective	**effect of**	Doxo drugs
compare	cytotoxic	**effects of**	displatin treatment

In Titles and Abstracts, we identified the role of 'of' in fixed terminology. In Introductions we find that fixed expressions have regular phraseologies beyond their internal components, possibly because there are simply more data for us to spot long range relations rather than because of any quality of Introduction sections. The term 'mechanism of action' appears to occur in a surprisingly delimited phraseological context: mechanism **of** action of {disease-related item} model {modalised 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 mutagenic
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...

A longer phraseology can also be seen in a common expression in Abstracts *treatment of*, which is now premodified by a combination of recurrent expressions in Introductions (we present one example of each):

{empirical problem or role} in/ for / by treatment ***of*** *{disease Y}*

...is a common clinical problem	**in the treatment of**	adult acute leukaemia
... expression... is induced	**by treatment of**	tumour cells with cAMP analogues.
... an alternative strategy	**for treatment of**	hepatoma...
...is... a promising candidate	**for the treatment of**	topical infections.

One particularly interesting term *'drug of choice'* (6 occurrences) collocates with '*in the treatment of*'. Even more striking is the level of reformulation of similar concepts for new drugs used in the longer phraseology: {treatment X} is a {new} drug (commonly) used in the treatment **of** {disease Y}:

aca C, a drug commonly used	**in the treatment of** breast cancer patients
APD a commonly used drug	**in the treatment of** cancer
Harris et al. suggest the drug of potential value used	**in the treatment of ..**.tumours.
(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

Of also introduces quantitative focus expressions in Introductions such as *a variety of*. The framework is involved in a longer phraseology: {biochemical process / entity or at times empirical process} is {used / empirical process} in (a) (wide) variety **of**- {treatment / disease related items):

Enzymes are involved	**in a variety of**	anticancer drugs
Both are inactivated	**in a variety of**	industrial drugs

Both are used as a solvent	**in a variety of**	industrial drugs
Methyl chloride is used	**in a variety of**	consumer drugs
Methylene is used	**in a variety of**	pharmaceutical applications

2.10 THAT$_1$ (Abstract salient word 7)

'That' as conjunction plays 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 used introduce *indicate*,

values	indicate	**that**
findings	indicated	**that**
results	indicate	**that**
information	indicated	**that**

The items *studies* and *results* also introduce *demonstrated*. A similar pattern is observed in Discussion sections. One difference with the Discussion section is the important rôle of *'that'* as relative pronoun in embedded clauses. *That* functions refers most often back to a specific chemical and establishing 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 emphasising the status of the knowledge structure (*allow prediction of experimental factors* ***that*** *underline our lack of understanding of these processes*). Such uses of *that* (and, indeed *who*) as relatives confirms Kretzenbacher's (1990) finding that embedded clauses are an important characteristic of Abstracts.

THAT$_2$ (Discussion salient word 1).

'That' is the most significant salient word in Discussion sections. The word is listed by *Wordlist* as one of the least salient words of the other rhetorical sections however, with the one interesting exception of Abstracts. In Discussion sections, *'that'* indicates the primary use of complement that-clauses which function as projections of research reports and facts (Halliday 1985:244). In terms of rhetorical function, *that*-clauses reformulate or evaluate results. *That*-clauses can be divided into four patterns in Discussion sections, in order of frequency of occurrence:

1) Research item + research process + hypotactic projections.
2) We / This study +research process + hypotactic projections.
3) Extraposed it + projections of modality.
4) Research item-embedded projections.

The first three lexical left-collocates of 'that' are all research processes involved in the first pattern (verb complement clauses: *suggest/s that, indicate that, show/n that*), but they have very different modalities associated with their subordinate clauses. The first example, *'suggest/s that'*, is introduced by an empirical measurement as subject, and the verb in the subordinate clause usually has some degree of modality or phase:

data	suggests **that**	reactive oxygen would be important
evidence	suggests **that**	simple sampling can be performed
the model data	suggest **that**	endothelin receptors might play a role
a number of observations	suggest **that**	MQ MT is unlikely to play a role in
lack of ...activity	suggests **that**	patients should be monitored

As a more affirmative expression, *'indicate/s that'* is introduced by deictic research process items as subjects and no modality in the subordinate clause:

These findings	**indicate that**	a cell has become committed to the.. lineage
These results	**indicate that**	the cell has been arrested early in.. development
The present study	**indicates that**	this parameter is highly correlated with
our data	**indicate that**	LIC is less immunogenic than other tumors

our data	**indicate that**	ras activation is an early event

Related to this structure, we find cleft noun complement clauses introduced by a limited type of empirical or research process subject:

The strength of this model {empirical)	is **that**
One drawback of such models {empirical)	is **that**
Another possibility {empirical)	is **that**
One disadvantage {empirical)	is **that**
The potential explanation {research)	is **that**
The main conclusion {research)	is **that**

The second pattern we find is syntactically the same as the first, except that the subject tends to be *'we'* or (depending on the verb) *'this study'* or the names of other researchers. The first most frequent pattern of this type *'showed that'* tends to entail more evaluation or negative results than its present tense counterpart 'show that'. Also unlike 'show that', it has 'we' and 'experiments' as possible subjects:

{Research item} *{Biochemical / Empirical process}*

Experiments	showed **that**	there was no homology in this region
we	showed **that**	there are no differences in drug uptake
studies	showed **that**	the compound was not an inhibitor
we	showed **that**	the parent compound was extensively metabolised
studies	showed **that**	active management was preferable

Another frequent expression, but which expresses a different phraseology is 'we conclude'. This time the subordinate clause deals with empirical explanation rather than quantification, and this tends to involve an evaluative modifier:

We conclude that	platinum orientation is not adequately represented
We conclude that	CTL and NK cells together play an important role

Christopher Gledhill (2000). *Collocations in Science Writing.*

We conclude that	ifosamine is well tolerated
We conclude that	MTT assay is suitable for assessing antiproliferative action
We conclude that	this in vitro behaviour is meaningful

Extraposed *it*-clauses (adjective complement clauses) permit the researchers to omit the research process subject of the main clause, generally involving almost obligatory modality in the complement clauses:

It is possible **that**	the bioavailability of BQ-123 might be different
It is possible **that**	abnormal gene product may be involved
It is possible **that**	P-glycoprotein may be responsible
It is possible **that**	serine phosphorylate could play some role
It is possible **that**	the MP modification could stabilise the... cuformation

In contrast, *it is likely that* involves modality, negative polarity, or some negation of a previous result:

it seems likely **that**	they missed the peak
it seems likely **that**	abnormal patterns affect...
it seems likely **that**	order and timing are not invariable
it seems likely **that**	cell counts were not carried in HMC100 p64
it seems likely **that**	... alterations did not reflect the PMN population

And in further contrast *it is clear that* is always used in opposition to previously negative results and introduced by adversative sentence adverbs:

Nonetheless	it is clear **that**	there are sex differences in metabolism
Nonetheless	it is clear **that**	cardiac effects are not dose limiting
Nonetheless	it is clear **that**	the glycoproteins were specifically induced
Although	it is clear **that**	TAA is not specifically induced
However	it is clear **that**	assignment is paramagnetically influenced

The fourth main pattern for *that* involves embedded noun phrase complements, and similarly demonstrates a modality projection between the noun and its embedded verb. One of the most frequent noun phrase complements is *'the fact that'*. The expression takes on a very specific rhetorical role, by first stating negative results and then by setting out an explanation:

The fact **that** *{negative empirical observation}* *{explanation}*

The fact that	[[this enhancement does **not** occur in females]]	implies that such oncogenes were not involved
The fact that	[[we **cannot** demonstrate this change]]	suggests that AIN causes different effects...
The fact that	[[the 150pp treated group was **not** killed earlier]]	might be due to weakness in the dose monitor
The fact that	[[2 MCR lines did **not** show higher activity]]	confirmed that these reagents were highly specific
The fact that	[[sequential accumulation of LOH was **not** observed]]	might be due to early monitoring

The expression <might be due to>, as seen in the examples above, is also related to the complex conjunction: <due to the fact **that**>. Here the writers reformulate some anomaly and then explain it, while the new explanation (which does not appear to be a reformulation of previous material) may constitute a research result in itself:

The **failure** of the two mechanisms could be	**due to the fact that** phenotypic substituents reach complex levels at low time intervals
These **discrepancies** were	**due to the fact that** antibodycolumns are rarely 100% efficient
The **ineffectiveness** of thiamine may be	**due to the fact that** thiamine has sizable groups present.
The **unexpectedly** high concordance is	**due to the fact that** multiple immuno processes are involved

We can see that *the fact that* appears to collocate across clause boundaries with the expression *due to* in the following example (it also consistently colligates with a negative expression): *The fact **that** we cannot demonstrate this degree may be due to insufficient sensitivity of our method.* Here we can

see reformulation at work, in that an anaphoric noun (an 'ownerless fact' in Francis's 1985 classification) introduces a subordinate clause which explains the fact. In the case of the last example, the negative result is embedded and the reformulation of the problem is presented as an explanation in the main clause. The idea that a subordinate clause 'explains' rather than sets results out is compatible with the semantics of the less frequent expression *'is explained by the fact that*.' Further proof of this is that we must thematicise the explanation in the last example or change the formulation to *'is the explanation of'* as in *'Insufficient sensitivity of our method [is the explanation of] the fact* ***that*** *we cannot demonstrate this degree'*. 'Insufficient sensitivity' can not be expressed as a negative result. This suggests that research processes are not valid explanations and are hence not permitted by the phraseology. The negative result / explanation pattern even extends beyond the level of the sentence, as can be seen from the following rather unique example (from JGM56D):

#1 We found that.. only anti B1 could mediate specific cytolysis.	#2 This is likely **due to the fact that** the difference is only one subclass.

The more frequent expression 'due to' reveals a regular pattern across sentence boundaries in other parts of the discussion subcorpus (#1 negative result or negative research process, #2 possible empirical explanation):

#1 Unfortunately we could not detect enzyme activity in crude extraction that converted cis ACHO8A to the transomer.	#2 This could be **due to** the instability of this activity in a cell-free system.
#1 The basis for this observed diffusion ... is not readily apparent.	#2 It may be **due to** inherent differences.
#1 However, control and treated levels of mutagenicity are not significantly different.	#2 This may be **due to** reduction in kinase levels.
#1 Levels of mutagenicity were not significantly different.	#2 This may be **due to** reduction of small intestinal glucoriadas.

These examples also reveal the important reformulating role of deictic 'this' which is discussed later. The phraseology of *The fact that* differs from alternative expressions, such as *the possibility that* where the embedded clause itself contains the modalised explanation (the main clause, not shown here, is usually an expansion of the hypothesis expressed in the embedded complement clause):

The possibility **that**	the hybrid cells might have differentiated
The possibility **that**	the chromosome changes might represent in vitro artefacts
The possibility **that**	B-chloro(...) may have contributed to...down regulation
The possibility **that**	this factor may contribute to the immuno-reversal
The possibility **that**	the higher p53 levels may be the result of unusually high

This expression forms a longer phraseological unit when it is introduced by clauses which express the modality of the proposition in terms of exclusion from or support for a research programme:

We cannot rule out	the possibility **that**
We should not rule out	the possibility **that**
Not only does this result eliminate	the possibility **that**
This does not exclude	the possibility **that**
These studies raise	the possibility **that**
These reports support	the possibility **that**

A similar phraseology accompanies the NP complement *'hypothesis that'* which is usually introduced by more positive results:

These data suggest	**the hypothesis that**	MGaa may be responsible
First evidence supports	**the hypothesis that**	...cell lines could be more resistant
Our observations support	**the hypothesis that**	MCChOH will occur only if deletion...
Our observations lend support to...	**the hypothesis that**	this might be the source of methylation
Our results are in agreement with	**the hypothesis that**	the promoting agent may resemble..

To summarise, we can divide the various *that*-complement clauses between those which evaluate results and those which reformulate and explain results as follows:

Evaluation:	**Reformulation**
suggest that (+modal)	indicate that
(empirical item) is that (+modal)	confirmed that
conclude that (+evaluation)	demonstrated that
showed that (+ neg. / modal)	show that (+/- neg.)
(we) reported that (+modal)	(we) reported that
it is possible that (+modal)	(we) found that (+quantification)
the possibility that (+ modal)	the observation that
the hypothesis that (+modal)	
Negative evaluation:	
it seems likely that (+neg.)	
(adversative) it is clear that	
the fact that (+ neg.)	
(neg.) due to the fact that	

Modality does not necessarily constitute evaluation: in the examples above we find that modality in most expressions accompanies other explicit markers of evaluation, such as evaluative modifiers. In many cases however modals have other uses, as discussed in the entry for 'may', below. Another interesting feature of the patterns is that some expressions maintain their collocational properties (such as negative polarity) in different syntactic patterns. In particular, the expression *'the fact that'* is the clearest case for arguing that the phrase has to be used where some negative result is present - whether that negative result in an embedded clause introduced by the expression, or in a preceding main clause (where the expression has to be converted into a clause linker *'due to the fact that'*) or even in a nearby sentence.

2.11 THERE$_1$ (Abstract salient word 4)

The significant use of *'there'* in Abstracts reveals a prevalence of simple impersonal extraposed clauses in this section of the article, most often expressing explicit evaluation about the shape of the research articles' results (up, down or no change):

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

The exclusive use of the past tense is in line with other expressions which express new results in the research article as a whole. These expressions typically precede highly significant items within the Abstracts subcorpus which deal with statistical direction or relation (*increased, decreased, interval, correlated*). The one or two exceptions to the pattern (qualitative empirical items) seem to highlight the preponderance of quantitative expressions elsewhere in Abstracts:

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

THERE$_2$ (Results salient word 7)

'There' has a role in existential clauses in Abstracts in the past tense evaluation of change in data. This is in line with the general finding

throughout the corpus that past tense or perfective aspect tend to correspond to current claims in the research article, whereas the present tense is used to express established fact or report past research. However, in Results sections the pattern moves to the present tense (*there is / are*) and tends to be embedded after NP or VP complement clauses. The most frequent pattern involves projection, where the main clause is generally a research process and introduces empirical observations with some degree of explicit evaluation:

Research process		*Evaluation*	*Empirical items*
it appears	**that there are**	considerable	differences (x10)
Topography confirmed	**that there are**	considerable	correlations
it is evident	**that there are**	important	differences
the fact	**that there are**	pronounced	correlations
we found	**that there is**	little	detectable activity
This indicates	**that there is**	no	redistribution
The observation	**that there is**	normal	overlap
Results show	**that there is**	some	protein development

The present tense is however replaced in the collocational framework *There was (_) evidence of / that*. The expression is used with negative evidence or some statement about more theoretical biochemical processes (but interestingly not without some modifier, and the simple expression *There was evidence of...* does not occur):

There was	no	evidence	of long term toxicity
There was	clear	evidence	of long term deterioration
There was	some	evidence	of tumor development
There was		evidence	of a decreasing risk
There was		evidence	that...viability was compromised
There was		evidence	for tumor development

What phraseological principle can be postulated to explain why tense corresponds with lexical choice in this way? One clue emerges in the phraseology of the extraposed existential expression *'there appeared to be'*. Researchers tend not to use this expression to signal data which are

problematic or which present a clear contrast (often preceded by *'Although'*). The verb *appeared* is consistently used as a hedging verb which collocates with negative data (in the right-collocates):

There (x16 occurrences)	appeared to be	low levels of expression
Although (x7) **there**	appeared to be	very few fibroblasts...
And (x8) **there**	appeared to be	slight correlation

There is a cluster of grammatical and lexical features which coincide with the negative *'There appeared to be no...*' pattern:

1. Existential 'there'.
2. Modality.
3. The use of the past tense.

Such clustering demonstrates that collocational processes extend beyond syntagmatic word-pairs and beyond the linear ordering of constituents. This may demonstrate that such a pattern exists as a marked form in relation to the more prevalent present tense pattern. The present tense pattern, with its thematicised research clause in Results sections, is a preferred way of presenting positive results, embedded within the modalised presentation of facts. The present tense is also used in a number of non-hedged demonstrative references in the present tense / that-clause pattern: *'This shows that... This indicates that'*). Generally speaking therefore, negative results serve as an aside or as a contrast with the main argument, while the present tense indicates that an argument is to be taken forward.

2.12 WAS_1 (Abstract salient word 6)

We have seen in our discussion above that the simple past is the preferred tense for presenting the research article's present methodology and results. Ironically, the present is used to introduce previous research. This appears to conflict with previous research (Hanania and Akhtar 1985) and Malcolm's (1987) distinction (past for generalisations, present for specific data). In the PSC we find that *'was'* generally reports the research article's {clinical} methodology and non-quantitative {empirical process} results. In Abstract sections, *'was'* can be seen to have a completely phraseological role to *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* + *that, for this, to suggest etc.* (contrast the

present tense with a negative in Abstracts, with the past tense usage in Results).

2) Extraposed *it* and *that*-clauses: *it is ...concluded, apparent, desirable, essential, important, possible, believed, expected, likely that...*followed by explanation.

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 a tumour's effect on the metabolism (*growth, weight, vasodilatation, expression*). As in Methods sections, *was* introduces passive participles which are often pre-modified by a technical (biochemical) adverb:

was	metabolically expressed
was	immunologically reacted
was	enzymatically deaminated
was	induced
was	carried

However, the majority of passives in the abstract are more empirical or research-oriented and resemble passives in Results sections: **was** + {research process} [ordered by frequency].... *observed, found, detected, determined, studied, seen, shown, investigated, demonstrated, performed, established, confirmed, compared*.

WAS$_2$ (Methods salient word 2)

Was / *were* have a relative consistent phraseology across the corpus, although in the expression *There was / there were* a different phraseology emerges in Results sections (as discussed above). The significance of *was* in Methods sections stems fairly straightforwardly from the prevalence of the passive in the past tense description of biochemical and empirical observations. Verbs used in the passive have very fixed collocational uses. A particularly frequent pattern emerges with *'detection'* which tends to be either <*carried out at*> {measurement item}' or *'accomplished* + {method}':

detection **was**	carried out at [X] mm (several instances)
detection **was**	accomplished using amplified PCR
detection **was**	accomplished using fluorescence differentials
detection **was**	accomplished using fluorescence techniques
detection **was**	carried out by the fluorescence model

When the verb is *'analyze'* the method is a statistical model:

the result **was**	analysed using the t-test
this [set of data] **was**	analysed using the general linear model
correlation of the assay group **was**	analysed using Student's t-test

When the verb is *'determined'* the method is a type of 'assay':

transferase activity **was**	determined using a commercially available immunoassay kit
the structure **was**	determined using a reverse-phase chromatographic assay
MAKIII expression **was**	then determined using the isotope-dilution assay
the reference range **was**	determined using 43 pharmokinetic assays

When the verb is *'performed'* the methodology can be a statistical or measurement-related item:

This analysis **was**	performed using exponentially growing cells
while our analysis **was**	performed using infrared spectroscopy
clinical determination of the title compound **was**	performed using an inverted microscope
baseline calculation **was**	performed using the t-test
cell line count **was**	performed using the Mann Whitney test

The repetitive nature of some of the methodological details in the corpus also reveals a number of fixed expressions (and even idiosyncratic idioms) involving *'was'*. The following examples are common to several different texts, although of course there is also much repetition within the same text:

the solvent was	removed under reduced pressure (x5 instances).
the solution was	run on the plates for the analysis (x5 instances).
the supernatant was	transferred to a new fraction (x6 instances, plus variants).
temperature was	maintained at (measurement) degrees C. (x7 instances plus variants)
the reference range for (drug X) was	(measurement x) nmol. (x5 instances)

The plural *'were'* tends to be used with plural biochemical entities (*mice, cells, controls* etc.) '{biochemical entities} ***were*** {clinical process verb} *by'*. Singular items on the other hand tend to have the following formulation: '{usually deictic} {empirical / research process} ***was*** {clinical / empirical process verb}. Thus singular and plural forms of the verb tend to coincide with different semantic verb classes.

2.13 WE_1 (Introduction salient word 8)

A rich set of alternative expressions emerges when the research article writers present their own previous or current research. The Introduction, together with Discussion sections, appears to be the privileged location for self-reference and overt justification of research goals. In many of the expressions referring to *'we'* there are time expressions or deictic references to the writing process. These appear to vary systematically according to the choice of verb and circumstantial adjuncts:

Here	**we** compare	production in sheep
Here	**we** compare	expression of gene alpha
Here	**we** compare	spectra

In this study	**we** examine	a combination of methods
in the present study	**we** examine	the activity of PKC (x2)
in a subsequent study	**we** examine	the incidence of protein

These time expressions may have a role in situating a present tense verb because the unmarked meaning of the present in articles is more usually to

report 'past' research or established facts. More frequently however, the researchers refer to themselves or to a generic audience (using *we*) in the perfective aspect (a form of present tense time reference but indicating 'recent past' research). This perfective pattern is complex but essentially contains the following recurrent elements: (time reference) {reference to this study / paper /report} **we** have {research process}:

<Recently>	**we have**	found that
Previously	**we have**	investigated whether
<In this paper>	**we have**	investigated reactive effects
In a previous paper	**we have**	investigated other protonated
In this study	**we have**	reported (x3)
In a previous report	**we have**	determined
In this report	**we have**	shown that
	We have <recently	been studying>
	We have previously	reported that mutant p53 causes
	We have recently	shown that (x3)
	We have previously	studied p53 expression
	We have previously	succeeded in catenating
	We have <in this study>	studied NAK cell susceptibility
	We have in this report	studied tumour-drug distribution
	We have in this paper	succeeded in establishing
	We have in this study	succeeded in establishing ph1-p

Generally speaking, when the research process is described by a metacomment (*investigation, report, study*), the sentence adverb as theme is placed sentence initially. When the verb is more technical or linked to specific empirical processes, the adverbial element is placed after the finite verb as a specifier of the technical verb. This is particularly clear with the verb *'report'* which is exclusively used in the simple present tense with

specific technical results or observed biochemical processes: **we** {time reference} {research process} {reference 'here'} (*that*) {results}:

we now report	that p53 overexpression is elevated in the presence of
we now report	that epoxyalcohol also inhibits
we now report	the results of our immunological studies
we report here	the results of a physical study
we report here	the results of our study
we report here	that 2DDP-subclones
we report	that growth in soft agar appears to involve.. substitution
we report	the synthesis of 3 substituted pyramidizole
we report	first isolation and characterisation
we report	characterisation of a new breast cancer cell line
we report	2 different approaches to synthesis

WE_2 (Discussion salient word 9)

The researchers' reference to *'we'* in Discussion sections is associated with cognitive research process (*we conclude, we believe, we consider*) whereas in the Introduction *we* tends to be used with 'research writing' processes to do with actions (*present, succeeded, compare*). This difference corresponds to our data on action-oriented *'to'* clauses which are more typical of Introductions than propositional *'that'* clauses (generally related to mental process verbs). In addition, *'we'* is subject of the following present perfect forms:

we have demonstrated, described, designed, detected, determined, developed, employed, established, examined, extended, found, identified, investigated, obtained, observed, noted, reported, shown, suggested, summarized, used.

Of these, *employed, extended* and *used* can be classed as clinical processes (on the basis of: *we have used clonogenic assays to quantify...*). More generally, writers tend to use 'cognitive' verbs when assessing negative results. Each verb however has a specific phraseology. For example, the result-specific *'we conclude that'* pattern technically rephrases an empirical result, while *'we believe that'* extrapolates and explains the outcome.

We conclude that… reformulation of results:

#1 A number of other approaches have addressed the assignment of change.	#2 **We** conclude that energy group effects are not overwhelming.
#1 T cells and NAK cells are essential for rejection.	#2 **We** conclude that CTL and NAK cells play an important role in the rejection of LAC-IL2 cells.
#1 The validation coefficient decreased from 6.3% to 6.4%	#2 **We** conclude that ... the dose expressed... does not contribute significantly.
#1 The result .. *did not reveal* a significant shift.	#2 **We** conclude that OS may affect the movement of PMNs.
#1 *Neither position* band was detected.	#2 **We** conclude that the glycoproteins.. are specifically recognised...

*We believe that…*evaluation of results:

#1 The cellular basis for this association is *unknown*,	#2 but **we** believe that comparing this in vivo... is meaningful.
#1 Even if methylene *does not interact* with hepatocyte...	#2 **we** believe that the magnitude is not sufficient.
#1 The reasons for the discrepancy are *not entirely clear*,	#2 but **we** believe that our technique of assessing transport... offers greater sensitivity.
#1 The relative LI's *did not differ* between methylene-exposed controls.	#2 **We** believe that methylene-chloride exposure did not provide a selective growth advantage.
#1 The role of the negative phosphate backbone... is *poorly characterized* at present.	#2 **We** believe that improved progress can be made to enhance understanding in areas such as chemical drug design.

Thus expressions introduced by *We conclude that* can (as the verb promises) stand as a summary of the main empirical observations. Expressions introduced by *We believe that* are not representative of the results but signal the perceived significance of the research in the eyes of the researchers.

3. The Phraseology of Research Article Sections

The data presented in the previous section set out the distribution of uses of single grammatical items as they are used in the research article. While most of the observations signal departures from predominant usage in the general language, certain features of language can be seen to vary relatively systematically from one grammatical item to the next. This was seen to particularly affect such general grammatical features such as verbal polarity, tense and complementation, clausal extraposition and projection and complex nominal modification. Grammatical items can also be seen to have consistent patterns in terms of semantic clusters and collocational sets and reveal consistent correlations between lexical or grammatical form and such discourse features as modality. Such data also suggest varying range of usage from one rhetorical section of the article to another. This section of the book explores this theme in more detail, by examining the specific role of grammatical items which are found to be statistically salient in one section of the article alone. I also set out here the statistics used to identify the grammatical items examined in the previous section (this data is also included in the Appendices).

3.1 Titles

There are only 2300 words in the PSC titles subcorpus. To study phraseology in Titles a larger control corpus was needed and so the Medline electronic database was searched for a diskfull of 572 titles relating to cancer (1 626 words) and, for comparison, their Abstracts were also analysed (58 332words) as detailed in section III.6. However, the items we analyse in the control corpus are determined by what is found to be salient in the PSC. The Wordlist programme gives the following data (in the same format as discussed in Section 2.6 above):

Table 11: Title salient grammatical items from the Wordlist program

Rank	**Word**	***PSC Titles* Freq.**	**% in subcorpus**	***PSC* Freq.**	**% in corpus**	**Chi sq.**	**Probability=**
12	OF	166	(7.6%)	21309	(4.3%)	59.3	0.000
60	FOR	110	(5.0%)	5224	(1.0%)	26.6	0.000
67	ON	24	(1.1%)	2182	(0.4%)	20.5	0.000
70	AND	99	(4.6%)	14610	(2.9%)	19.7	0.000
134	IN	91	(4.2%)	14349	(2.9%)	12.9	0.000

A Wordlist comparison of the Medline Titles corpus and their corresponding Abstracts reveals similar data for grammatical items: *of, on, and, in, by, via, its*, together with the marginally grammatical *self* (in relation to self-analysis techniques for breast cancer). Most of these items have been analysed above, and only the item *on* remains.

3.11 TITLE salient word 3: On

'On' occurs in expressions that are either the topic of research or the application of a specific empirical process. A limited set of items introduce *on*, and its typical left-collocates have been listed under *'of'* above (disease related items):

{Research processes}:		*{Empirical processes}*
a retrospective study	**on**	effect
Basic study	**on**	influence
Clinical study	**on**	impact

In Titles *'on'* is also a key element in fixed modifying expressions which add embedded information about methodology, as in {research process 1} based **on** {research process 2 / clinical process}:

{Empirical process}		*{Research process}*
design for pilot studies	based **on**	lab data
lymphatic studies	based **on**	a clinicopathological study

flow in carcinoma	based **on**	anatomic manner of extension
design methodology	based **on**	NMR combined spectroscopy

On is less involved in complex nominals than *'of'* and *'for'*. As mentioned in our discussion of *and* and *of* (both Title-salient items) prepositions such as *on* are largely determined by the widespread use of lexical items such as *effect*. The collocational relation between *effect* and *on* can be seen to operate regardless of complement or modifier roles, especially when the item *'effect'* is seen to govern a prepositional complement phrase:

1 The effect of surgical intervention and neck cancer **on** whole salivary flow. (Modifier of *effect*)
2 Blood transfusion does not have adverse effect **on** survival after operations for colorectal cancer. A pilot study. (Complement of *effect*).

In #1, the prepositional phrase can be inserted before the presumed complement phrase *introduced by *of*). The proximity of *effect* and *on* in #2 suggests that *'on'* belongs to a complement phrase (if no other material can intervene in that position), in which case *after* is candidate for introducing a modifier. In either case, if *'effect'* is seen to introduce *'on'* then a collocational relation appears to be valid across phrase boundaries.

3.2 Abstracts

There are 29 136 words in the PSC Abstracts subcorpus. The *Wordlist* data reveal the following salient words:

Table 12: Abstract salient grammatical items from the Wordlist program

RANK	***WORD***	***PSC Abstracts Freq***	***% in subcorpus***	***PSC Freq.***	***% in corpus***	***Chi sq***	***Probability***
31	BUT	67	(0.2%)	663	(0.1%)	18.1	0.000
43	THESE	119	(0.4%)	1399	(0.3%)	15.3	0.000
79	OF	1367	(4.7%)	21309	(4.3%)	11.8	0.001
198	THERE	40	(0.1%)	444		6.5	0.011
203	IN	912	(3.1%)	14349	(2.9%)	6.3	0.012
267	WAS	365	(1.3%)	6271	(1.2%)	5.0	0.020
299	THAT	227	(0.8%)	3357	(0.7%)	4.5	0.034
329	DID	34	(0.1%)	395		4.3	0.037
334	WHO	14		129		4.2	0.040
378	BOTH	55	(0.2%)	713	(0.1%)	3.7	0.055

The salient lexical items of Abstracts 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*). 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 which report past research (*studied, suggest*). This tendency is borne out by the phraseology, as we have seen above for items such as *of*, *there, in, was, that, did*, The following four salient items are uniquely significant in Abstracts sections, and confirm the general tendency for embedded expansion (in clauses and phrases) and quantitative reporting.

3.21 ABSTRACT salient word 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. Positive results are announced in a first clause and then qualified. 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 (items include *resistance, efficacy, immune response*). In Results sections on the other hand, we find that the tendency is to explain negative results using adversatives which introduce hypotactic clauses rather than co-ordinating conjunctions (***however****...X did not correspond,* ***although*** *this did not result in...*). As we have seen above, in Abstracts report and quantify negative data whereas Results expand on and qualify them.

3.22 ABSTRACT salient word 2: These

'This' serves as a determiner (in rephrasing, or reformulation) or as a deicitic pronoun to refocus information from one clause to the next. This function is shared by Discussion sections and a more detailed analysis is seen in our discussion of 'this' below. We note here that 'these' in Abstracts differs from 'this' in that almost half of the occurrences of *these* are as pronouns introduced by *of*, while 'this' is mostly a determiner. The anaphoric referents of *these* tend to be very specific disease-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 with similar (but infrequent) patterns for *of* (see previous section). Abstracts therefore tend to favour the use of deictic encapsulation (pointing to single items) as opposed to reformulation (a process seen in Discussion sections, where *this* and *these* are determiners of longer noun phrases rather than single pronouns). The high significance of

these (according to Appendix C2) here also coincides with Nwogu and Bloor's (1991) observation that abstracts tend to employ simple thematic progression, linearly converting rheme to theme.

3.23 ABSTRACT salient word 9: Who

The relative pronoun *who* is prime evidence of embedding in Abstracts (also seen in the pronominal use of *that*). *Who* refers to the only participants other than the researchers (*we*) mentioned in the corpus: *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 (a passive expression), they receive them (*who receive carboplasmin, receive Doxo, receive doxorubicin*). This may be legal requirement or a deliberate euphemistic avoidance (unlike mice, patients must be willing recipients of drugs) – although the consistency of the expression in the corpus and the fact that science writers are not aware of such conventions suggest that we are dealing with a very dominant scientific 'voice'. This is also quite a clear example of the way phraseology helps to shape a specific view of 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 *active physiological management* and *administration*. The phraseology of the term *management* (the 46th most frequent term in the corpus) allows us to establish its meaning within the corpus not only as very different to 'organisation of personnel' but

as part of a larger, 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 / were fed / were 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.

3.24 ABSTRACT salient word 10: Both

'Both' signals a noun group complex, another possible characteristic of 'compaction' in Abstracts. In many of the cases where *'both'* is used as a linking conjunction, it is a redundant signal of a following conjunction. The following sentence is typical:

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

As mentioned in our discussion of *and* above, *'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

Both appears to signal a paradoxical relationship between two terms at extreme ends of a scale, establishing at the same time the limits of the scale (short range: from *mice* to *rats,* or long range: from *normal* to *tumor cells*). By using such expressions in Abstracts, the writers signal a broad and inclusive data set to be compared in the research article.

3.3 Introductions

The PSC introductions subcorpus contains 59 724 words. The Wordlist comparison with the PSC gives the following data:

Table 13: Introduction salient grammatical items from the Wordlist program

RANK	**WORD**	**Abstracts Freq.**	**% in subcorpus**	**PSC Freq.**	**% in whole corpus**	**Chi sq.**	**Probability=**
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		73.7	0.000
15	CAN	120	(0.2%)	468		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%)	21309	(4.3%)	41.4	0.000
32	TO	1233	(2.1%)	8631	(1.7%)	36.6	0.000

The phraseology of these items indicates a general tendency for extraposed projections (clauses of action and hypothesis), the relational expression of technical facts, the reporting of previous research and the present signaling of research goals. The lexical properties of Introductions are considerably more complex that those of Titles and Abstracts and, generally speaking, the

phraseology of Introductions is distinctly unlike that of the rest of the research article.

3.31 INTRODUCTION salient word 1: been.

'Been' is used in two types of perfective passive construction which have been identified as typical in the reporting genre of Introductions (Salager-Meyer 1992). We have seen many of the phraseological properties of the perfective in our discussion of *have* (above). The passive perfect appears to polarise around a semantic difference between research process verbs introduced by a biochemical / empirical subject and verbs which indicate a new or prevailing theoretical model in extraposed clauses:

1) {biochemical entity or research process} (has / have) **been** {research process verb} *in order of frequency >10*: *reported, shown, demonstrated, found, observed, identified, studied, described, obtained, published, conducted, detected, investigated*. However, this 'report' pattern also involves three empirical process verbs: *used, implicated, associated*.

2) it has **been** (*in order of frequency >10*: *shown, suggested, proposed, established, postulated, concluded*) that. These are also research process verbs as we have defined them above, but they also tend to be mental or verbal processes (Halliday's terms) and refer more to the research activity of the discourse community than to that of the authors. The whole pattern is termed a 'research utterance'.

The verb *'shown'* appears in both lists, and I claim below that it has a different distribution to other verbs. However, the most significant right-collocate of *been* with 40 occurrences is *reported* in the following phraseology: (biochemical process} has/have **been** reported to (+ quantification clause}:

p53 gene resistance	has **been** reported	to be very frequent
drug resistance	has **been** reported	to be different in 2 case studies
antigen mechanisms	has **been** reported	to be frequently carcinogenic
the LOH mechanism	has **been** reported	to cause significant immunological damage
S-transferases	has **been** reported	to produce metastasis in several species

A less frequent but similar phraseology involves *reported in* (+quantification phrase}:

gene inactivation	**has been reported in**	a number of cancers
MP substitution	**has been reported in**	a high percentage of carcinomas
LOH from 18q	**has been reported in**	several human cancers
low effects of inhibition	**has been reported in**	many tissues
drug resistance	**has been reported in**	mammals treated with PIMO

This appears to be a typical pattern for other research process verbs (*observed, described, detected*). When we analyse the empirical / relational process *associated* in the same global pattern, the expression relates tell-tale signs of cancer to causes: {biochemical process} have been associated *with* {cancer Y}:

Retroviruses	has/ have **been** associated with	hepatic cancer
Ras gene	has/ have **been** associated with	specific neoplasia
high doses of toxin	has/ have **been** associated with	gastrointestinal bleeding
mutation in these genes	has/ have **been** associated with	haemic neoplasms
its effects on human health	has/ have **been** associated with	the occurrence of cancer

While this may appear to be unremarkable, it has to be remembered that quantification is a possible pattern with *associated* but is simply not used. A similar pattern is seen with *implicated* except that the pattern is: {biochemical process} have **been** *implicated in* {disease-related process Y} and the disease-related item is more specific than in the *associated with* pattern:

...have **been** implicated in...	regulating cell differentiation
...have **been** implicated in...	in the development of cancer
...have **been** implicated in...	the t-programming process

The third exceptional empirical report in the first pattern has a unique phraseology, involving a statement about a general research model or technique as subject:

This model	has	**been**	(widely)	used...
animal models...	which have	**been**		utilized....
This type of assay	has	**been**		used...
the macrolide technique	has	**been**		used...
A cross-characterisation technique	has	**been**		utilized....

Utilized is mostly interchangeable with *used* but is less frequent:

... have **been** used/ utilized	to study / evaluate / prepare... {biochemical X}
... have **been** used	for other TCNQ derivatives
... have **been** utilized	for the commercial production of citric acid
... have **been** used	as a guide in the primary study
... have **been** utilized	as chiral auxiliaries in a variety of assays

The difference between the two verbs is that *in* only follows *utilized* :

... have been **utilized in**	industrial settings
... have been **utilized in**	combination chemotherapy
... have been **utilized in**	a recent synthesis
... have been **utilized in**	the delivery of amines
... have been **utilized in**	cancer therapy

Such differences imply an extra level of phraseology available for this expression, and may indicate the effects of American English on the general phraseology of the corpus.

The clauses introduced by the second major pattern (extraposed + research utterance) have a less technical semantic scope than those in the first and generally express some empirical relational clause (*X is associated with / involved with Y*). The projected clause is a past result framed in terms of a new (present tense) research direction (the following examples are listed in order of right-collocate frequency):

it has been proposed that	this transformation involves DNA damage
it has been established that	they are reactive with the extracelluar domain of p185
it has been postulated that	the mitogenic effect of estrogens are mediated
it has been concluded that	MP substitution is a significant tumorigenic factor.
it has been suggested that	thyamine is involved in the development of prostatic cancer.

I suggested above that collocational patterns are not due solely to the grammatical preferences of lexical elements (in this case verbs) but to a general semantic 'meaning' that the collocational framework embodies. A clear example of this can be seen with *'show'*. Since *'show'* appears to fit semantically into several categories of verb (empirical and research-oriented) it is perhaps no accident that it is the sole verb to be used in both the passive perfect 'reporting' pattern and the extraposed 'research utterance' pattern. Furthermore, its use does not quite coincide with other verbs in terms of phraseology and lexical collocation. In the first pattern (24 instances), the expression introduces non-finite clauses in the same way as the verb *report*. In this case, however, the clause does not present quantitative results (found exclusively after *has been reported to*) but more qualitative findings:

the disease	**has been shown to**	have considerable resistance to
TNF alpha	**has been shown to**	efficiently deliver the toxicity of ricin
a structural analogue of histidine	**has been shown to**	provoke an immune response
Quercetin, a lipoxygenase inhibitor	**has been shown to**	exhibit antitumour activity in vitro
encapsulation of dXR...	**has been shown to**	act as an in vitro inducer

The extraposed pattern for *show* is similar to other verbs such as *establish*, which introduce an explanation rather than a specific quantifiable result. The difference with other verbs lies in the choice of clause complex, and *show* is used almost exclusively in thematically prominent subordinate clauses introduced by *Although*:

Although it has been shown that	the murine p53 used in all of these studies was mutated, its mechanisms are not fully understood.
Although it has been shown that	p53 gene constructs with many different point mutations, the gene responsible for the two cancers has not been identified.
Although it has been shown that	the hepatocytes are critical to the survival of the tumor, ... no correlation has been previously determined...
Although it has been shown that	the cells that mediate cancer induced GVHD, structural studies of the enzymes have yet to be published.

Show is thus used almost exclusively to present contradictory evidence which has not yet been published. These sentences are a clear case of consistency of use, and demonstrate that collocational behaviour extends beyond the level of the clause. We can see that the expression *'it has been shown that'* has a specific phraseology but is not incompatible with the other research utterances. It plays a marginally different role to these expressions, and writers choose it to distance themselves from the possibly more subjective 'cognitive' verbs of the same phraseology. Why should the extraposed *show + that* clause be limited to signaling gaps in the research record? It may be that the semantics of the verb *'show'* are sufficiently vague and non-emphatic (as opposed to *proposed, concluded, established, suggested*). This allows the writer(s) to suggest a framework in which the wider discourse community has no agreed fixed position on previous findings (neither proven nor rejected).

3.32 INTRODUCTION-salient word 2: Has.

As with *'have'* and *'been'*, *'has'* plays a key role in the phraseology of report, taxonomy and evaluation. *'Has been'* accounts for 60% (188/284) of the instances of *'has'*, and this usage is detailed above. The remaining phrases using this item are collocational frameworks with *'of'*: *have the _ of* in which the whole expression functions as an attributive relational process:

has	the advantage of
has	the benefit of

has	the characteristic of

There are also a number of instances of impersonal reporting in which the phraseological pattern is: {clinical approach or technique) **has** received {quantification of research process} 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 potential cure for cancer
C1350	**has** received particular attention as a possible source of metabolic data.

As seen elsewhere in the corpus, the relational or possessive use of 'has' also involves overt evaluation:

the inhibitor	**has**	**a**	profound	effect on its structure
the factor	**has**	**a**	peak	incidence between...
the disease	**has**	**a**	broad	spectrum of clinical indications

3.33 INTRODUCTION salient word 5: such.

The expression *'such as'* is a discourse marker reformulating items in a taxonomic way. The most frequent reformulations are of biochemical processes (*agents, enzymes* and *tumours*) where the reformulation demonstrates the conventional notation or chemical nomenclature for the superordinate chemical type:

antitumour	agents	**such** as	NMU
alkylating	agents	**such** as	BCNV
carcinogenic	agents	**such** as	nitromidazoles
other	agents	**such** as	TCPOB-08

use of hormonal	enzymes	**such** as	dismutase

several DNA	enzymes	**such** as	exonuclease
metabolic	enzymes	**such** as	transferase
detoxifying	enzymes	**such** as	acetates

tumors	**such** as	Wilm's melanoma
tumors	**such** as	maleic myeloma
tumors	**such** as	the adenocarcinoma
tumors	**such** as	MCF-7

The reformulation appears to be bi-directional: the first item can be a new item, while the complex preposition *'such as'* introduces a reference to a previously mentioned specific item. In this case, the textual function 'given' or 'new' does not determine word order, the phraseology (superordinate) such as (hyponym) remains the same. The 'new superordinate / given hyponym' reading of this pattern is not listed for this expression by the Cobuild dictionary, and it is plausible that particular uses of set expressions like this undergo slight shifts of use in technical writing. What is clear, however, is the function of rephrasing (reformulation) which confirms that this is a fundamental mechanism in report writing and explanation in Introductions. This also occurs in a slightly different form to Discussions: reformulation in Abstracts and Introductions can be seen to 'refocus' single items, while Discussions sections reformulate items as more generic terms.

3.34 INTRODUCTION salient word 6: can.

'Can' expresses potential empirical procedures or biochemical processes. The verb essentially signals a reduced form of claim. Two patterns emerge, either in research oriented passive constructions or in active technical expressions:

2) {General clinical or empirical process} can be {research / empirical process }:

alterations	**can**	be	prepared	applied
variants	**can**	be	deciphered	prevented
ideas	**can**	be	correlated	determined
methods	**can**	be	considered	classified

therapies	**can**	be	attributed	derived
products	**can**	be	obtained	

Some technical biochemical processes are also used in this expression: *transmitted, modulated, coupled, induced*.

3) {Specific biochemical process / item} **can** {technical biochemical process}:

gene products	**can**	dimerize
cytokines	**can**	flip
IL-2	**can**	hydrolyse
differentiated cells	**can**	induce
gingivalis	**can**	undergo malignant transformation
DNA	**can**	metabolise
PMEA	**can**	inhibit

In Introductions, at least, the passive is not used to express clinical or technical biochemical processes. This trend is reversed in Methods sections, as we have seen for *was* / *were* above.

3.45 INTRODUCTION salient word 7: It.

Most of the uses of 'it' have been described in the discussion of *'it is'* and *'it has been'* + {research process} above. While the present tense is the preferred tense in Introductions, with the verbs *found*, *thought (x3), reasoned, reported, shown* the extraposed passive is expressed in the past tense:

it was also	found that	the polymer was not stable
it was	found that	it causes higher overall cell counts
it was	found that	although stability outside the cell...

'It' is the most Cobuild-salient word in the corpus. The Astec 'Common' program shows that in relative frequency (not actual frequency), it is nearly five times more likely to occur in the Cobuild corpus than in the PSC (the ratio is 20: 112 per 1000) and this would indicate that extraposed clauses are a prototypical characteristic of Introductions rather than the rest of corpus. Extraposed active clauses (in *that*) are however overtaken in Introductions by the use of non-finite extraposed *to*-clauses, such as evaluative research utterances (*it is essential to* etc.) and *it would be worthwhile to*. Such action-oriented phrases are described below.

3.36 INTRODUCTION salient word 10: To.

Generally speaking, the prevalence of *to* in Introductions is indicative of a preference for action-oriented clauses as opposed to cognitive 'mental' process clauses. Such a distinction was first observed from concordance data by Johns and King (1993) in the general language. In the PSC, *'it is important to'* and *'have been reported to* are followed by specific findings or empirical events. This can be contrasted with present tense or modal expressions such as: *it appears that*, and *it would seem that* which tend to introduce hypotheses and explanations (as seen under *been* above: *to* clauses such as *has been shown to* are more frequent than *has been shown that*). The most frequent use of *'to'* as complementizer is in projecting cleft clauses which formulate the aims of the research paper, a key expression in Introductions sections. We have already seen *'This aim of this study was to'* in our discussion of *of*, however the variety of expression we find with < *was to* > goes well beyond this simple formulation:

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 study's 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.
(Exception: The purpose of the current report	was **to** generate and trap...)

The only permanent elements of the phraseology here are the grammatical items *'was to'*, and the semantics of the surrounding clusters is highly consistent: {research goal} *was to* {research process verb}. The only exception to this seems to be where the aim is to act in a specific methodology, for example the clinical process *'generate and trap'*. This may seem unsurprising, but the important point about phraseology is that perfectly plausible alternatives such as *'to generate and trap*' are not equally as prevalent as the 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. In the case of Introductions, goals are presented as global research rather than the specific empirical or clinical processes. A possible corollary is that what would be free or restricted collocation in the general language becomes fixed either one way or another in the specific language because of such overriding rhetorical constraints.

However, this does not exhaust the role of *to* as complementizer in noun group projections in other salient expressions in Introductions. One particularly regular projecting clause takes the 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

In some cases, adjective complement clauses reflect more typical verb complement patterns. *'Able to'*, for example can have animate subjects {the researchers} with the following pattern: (we are/were) able to {research process}:

we were	able **to**	compare the patterns
we are	able **to**	confirm that...
if we were	able **to**	design an interim system
we are not yet	able **to**	give a definitive statement
In 16 cases we were	able **to**	identify the structural defects

or inanimate biochemical subjects with the following pattern: {biochemical process / entity} (be) able to {biochemical process}:

agents that	are able **to**	down regulate
gangliosides	are able **to**	function as
human IL2	is not able **to**	induce an immune response
the most potent of these	is not able **to**	maintain cAK III
The...analogous tumor	was also able **to**	metastasize.

This phraseological distinction {research oriented / biochemical oriented} is also strikingly reflected in the tense patterns of one 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 (and can be seen to be used in reporting of results):

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

This appears to confirm our findings elsewhere that tense and aspect play a role in phraseology (we see elsewhere that it does for *is / was / have been*). Rather than representing a stance in relation to past and present (current) research, the past tense appears to correspond to research-oriented observations (relating to the overt mental or verbal activities of the researchers) while the present corresponds to biochemical and empirical observations (covert activity on the part of the researchers).

I have mentioned above that projected *'to*-clauses' (such as the very frequent *have been found to, designed to*) are characteristic of Introductions while projected 'that-clauses' (*The possibility that, it has been found that*) become are preferred in Abstracts and Discussions. This may reflect an increased use of indirect grammatical metaphor later on in the text. In Introductions, for example mental research processes (in the passive) project explanatory clauses impersonally:

cells	are	known **to**	bind p53
chemicals	are	known **to**	cause embryotoxicity
enzymes	are	known **to**	inhibit hepatic MFO activity
hydrolysis	are	known **to**	proceed via a 2-step reaction
proteins	are	known **to**	repair the 6-0 methylguanin

If we look at the long range phraseology of the most frequent of these expressions 'appears to' we see that 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 of 'appears to' seems to be linked not with 'hedging' of assertions, as one might expect, but with signalling contradiction, tied in with negative subordinate clauses. It is also worth noting 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 lexico-grammatical choices are available, with the final mechanism of thematic choice and word order left to textual considerations.

Finally, the prepositional use of *'to'* accounts for only half of its occurrences in Introductions whereas it becomes prevalent in Methods sections. In particular we note its use in the adjunct: *according to* + research model (*in vitro criteria, soliton theory, the theory of Knudson (1985), the mechanism we put forth, tumor histology (Palmer et al. 1988)*), phrasal verbs, as with the very frequent *compared to* + biochemical process, and complements of biochemical nominals which take -*to*-, such as the frequent *'resistance to chemotherapy'*. A longer phraseological unit emerges with the nominal {empirical process} {emprical premodifier} *exposure to* {biochemical entity}:

(drug X) was increased following short term	exposure **to** TNF and other solvents
(drug X) undergoes induction involving	exposure **to** high concentrations of TNF
Studies have demonstrated permeability following exposure **to** non-toxic doses	
industrial	exposure **to** methylene chloride
human	exposure **to** higher concentrations
occupational	exposure **to** benzocaine

Other nominal constructions normally use *'to'* phrases as a comparator, very often involving *'cells'* and another biochemical, often a reagent *'growth factor'*:

responses	of cells **to**	a wide variety of mitogenic growth factors
resistance	of cells **to**	growth factors
susceptibility	of cells **to**	hormones in growth factor

responsiveness	of cells **to**	oestrogens
similarity	of cells **to**	the antibody

3.4 METHODS sections

The PSC Methods subcorpus contains 137161 words. The Wordlist comparison with the PSC gives the following data:

Table 14: Methods salient grammatical items from the Wordlist program

RANK	***WORD***	***PSC Methods Freq***	***% in subcorpus***	***PSC Freq***	***% in whole corpus***	***Chi sq.***	***Probability=***
1	WERE	2795	(2.0%)	5162	(1.0%)	876.5	0.000
3	WAS	2877	(2.1%)	6146	(1.2%)	576.7	0.000
18	THEN	282	(0.2%)	420		142.9	0.000
20	AT	1324	(1.0%)	3287	(0.7%)	140.3	0.000
25	FOR	1919	(1.4%)	5224	(1.0%)	120.1	0.000
30	EACH	323	(0.2%)	595	(0.1%)	100.2	0.000
44	AND	4633	(3.4%)	14610	(2.9%)	74.3	0.000
82	FROM	1048	(0.8%)	2982	(0.6%)	47.2	0.000
139	AFTER	431	(0.3%)	1139	(0.2%)	32.0	0.000
260	WITH	1711	(1.2%)	5543	(1.1%)	17.8	0.000

The language of this section is adapted to express very specific sets of instructions, accompanied by a marked lack of subordination and often resulting in the progressive use of shorthand abbreviations in experimental sections. The expressions to be found in this section are thus highly regular and presumably help the 'indexical' reading of the text.

3.41 METHODS salient word 1: Were.

As with *'been'* in Introduction sections, 'were' is indicative of the passive. But whereas passives elsewhere in the corpus tend to be research oriented (*'have been identified'*, etc.) here the past passive (which is largely unique to the Methods section) is clinically or empirically oriented, involving sometimes highly technical verbs. This contradicts Hanania and Akhtar

(1985) who found that the passive in Methods was found to be frequently present tense (*is identified, has been identified*). Conversely Heslot (1982) and Wingard (1981) found that the simple past was prevalent in Methods sections, which also appears to be contradicted in this corpus. In the literature, passive expressions in science writing have been characterised as a novel relationship between subject and verb (Sager et al. 1980, Heslot 1982, Hanania and Akhtar 1985, Swales 1990). It can be seen that grammatical subjects correspond consistently with either clinical or empirical verbs (with some exceptional cross-over):

anaerobes	**were**	(empirical) enumerated
analyses	**were**	(clinical) carried out, performed, prepared
animals	**were**	(clinical) allowed food, given food, housed in quarantine randomly assigned / allocated a cage, killed,sacrificed
cells	**were**	(clinical) collected, cultured, fixed, grown, incubated, maintained, plated, seeded, sonicated, subcloned, treated, trypsinised, washed (empirical) counted
compounds	**were**	(clinical) separated, dissolved, heated, dissolved, obtained, prepared, combined
concentrations	**were**	(clinical) optimised, added, adjusted, maintained (empirical) achieved
data	**were**	(empirical) pooled , expressed, obtained (research) analysed, considered
mice / rats	**were**	(clinical) bled and killed, exposed to, fed, given killed, observed, obtained, raised, treated, weighed
patients	**were**	(empirical) asked for their consent, entered at many intervals, excluded from the study, followed until death, (clinical) treated at dose level
samples	**were**	(clinical) collected, obtained, run at x%, centrifuged (empirical) counted
tissues	**were**	(clinical) fixed, homogenized

However, patterns of the passive can perhaps be more usefully sorted according to the elements which follow the passivised verb, which are for the most part prepositional modifiers (adjuncts). We see later that these can be further sorted by verbal process. I term such sorting of phraseology from one pattern to a sub-pattern 'collocational cascade' because this is the effect of the listing on the page. Thus the most frequent pattern for the passive is: {biochemical entity} were {clinical process} by {biochemical entity} (detailed in a later section). Setting out other passive + preposition patterns we find that the collocational cascade takes on a further 'step' since each passive then has specific (but consistent) element with a sense of instrument / medium:

were	analysed by	log rank test
were	analysed by	ANOVA test
were	analysed by	using analysis of variance
were	determined by	TLC scanner
were	determined by	liquid scintillon counting
were	determined by	the method of Chadwick et al.
were	determined by	means of a Student's t-test
were	determined by	the HPLC method
were	killed by	cervical dislocation
were	killed by	exsanguination
were	killed by	CO2 anaesthesia
were	killed by	CO2 asphyxiation
were	obtained by	measuring the fluorescence{clinical procedure}
were	obtained by	using a 1.5 mm diameter cork borer
were	obtained by	retro-orbital bleeding of mice
were	obtained by	injecting 3x105 cells into both flanks

were	prepared by	the reverse evaporation method
were	prepared by	the film method of Skoza et al.
were	prepared by	protein precipitation with acetonite
were	prepared by	dilution of the liposome dispersions

Such a use of *by* for the medium of the sentence rather than the agent changes our stereotypical view of the passive (in which *by* signals a grammatical agent: *prepared by the scientists* etc.). In a collocational framework with 'for' (a Methods salient word) the passive construction is empirically oriented rather than clinical:

were	analysed for	{observable item} hormone traces
were	analysed for	significance
were	calculated for	antibody depletion
were	calculated for	luteinizing hormone count
were	eligible for	{study} the present study
were	eligible for	this study
were	examined for	{disease-related item} visceral defects
were	examined for	malfunctions
were	examined for	external defects
were	used for	{research process} observation
were	used for	evaluation of patients
were	used for	the experiments

With *'at'* (another Methods salient word) the passive construction is used to express some measurement together with clinical process verbs. As with the patterns above, the collocational cascade only has one step in this pattern since the phraseological possibilities for circumstantial elements are limited to times/ temperatures:

{Clinical process}

were	collected	at	appropriate time levels
were	collected	at	77 minute intervals
were	collected	at	1 minute intervals
were	incubated	at	37 degrees C
were	stood	at	room temperature
were	performed	at	37 degrees C
were	repeated	at	room temperature.

The overall picture seems to be that we can usefully categorise certain passive constructions by the types of prepositions that are used to signal adjuncts in these expressions. These are of course mediated by the specific phraseology of passivised verbs, and these verbs and their subjects and adjuncts can in the majority of cases be classified semantically and regularly subclassified by verbal process. However, there are also various choices of expression for the same process. For example several idioms are used to express the (legally obligatory) destruction of animals. Here are the possibilities in decreasing order of frequency (subjects include in order of frequency: *animals, mice, rats, rabbits, pigs, monkeys, dogs* and *'control groups'*):

{animals)	**were**	killed	by cervical dislocation
{animals)	**were**	sacrificed	by severing the dorsal aorta
{animals)	**were**	euthanized	after 82 weeks
{animals)	**were**	necrotized	by CO2 asphyxiation

3.42 METHODS salient word 3: At.

Prepositions such as *by* and *at* have virtually only one use in the cancer research article as opposed to a wide range of use in the general language. '*At'* signals empirical measurement or quantification, either of temperature, duration or increments of time. *'At'* is necessary after a wide range of passivised clinical process verbs as we have seen with *'was / were'*, or within the collocational framework of *'for* (x hours) *at* (temperature x):

centrifuged	**at** 12 000 rpm

eluted	**at** a flow rate of
heated	**at** room temperature
incubated	**at** room temperature
measured	**at** 400mm

As stated above many of these are repeated several times within the same text, and listed in the methods section so that certain phrases achieve the statistical status of idioms. Here is just one example of many, although we can claim that this is unique in that it involves a triple collocational framework with an inverted temperature / time expression (as compared with the expressions above): *was* (stirred) *at* (temp.) *for* (time.) *until* (empirical / clinical process item}:

was stirred **at** 20 degrees C. for 40 min.	until DNA extraction
	until processed
	until assayed
	until analysed

There are also a number of idiomatic uses of *'at'*, for example the expression 'at risk' in apposition to either *tumors / carcinomas* or *animals / mice*. The lexical phrase *'at least'* is perhaps the only exception to this general modifier pattern, although it also fits into the broader expression of 'measurement':

total of	**at** least 15 000 nuclei per sample
expectancy of	**at** least 60% a load
model cohort of	**at** least 3 patients
based on	**at** least 4 tumours
performed on	**at** least 2 separate occasions

The 'location' meaning of *'at'* is rare in the corpus, although we find instances such as: *unidentifiable numbers are placed at the bottom of the scale*.

3.43 METHODS salient word 4: Then.

We have seen above that the number of uses listed in Cobuild dictionary for certain words is usually highly restricted in the PSC. Although *then* is an important feature of narrative in English, there is simply no need for argumentation in this section of the research article and despite being a very significantly 'Cobuild-salient' item, *'then'* functions here in a restricted way (it corresponds to 1 out of 10 possibilities in Cobuild (1995 2nd edition): as a time-specifier before passivised verbs to signal a subsequent incremental step in the methodology. The most fixed phraseology involves an idiomatic expression *'the solution was added dropwise and the suspension was then heated'* (x4 instances). The following clinical verbs are most frequently used in this construction:

the solution was cooled	and **then**	added
the supernatant was internalized	and **then**	extracted
fifteen slides were exposed	and **then**	incubated
the frozen cells were thawed	and **then**	transferred
the mixture was filtered	and **then**	washed

3.44 METHODS salient word 6: Each

The determiner *'each'* is evidence of deictic refocusing, in which the researchers emphasise the distribution and repetition of a series of clinical processes:

{Empirical quantification : application of a dose}

verified	at **each** dose level
entered	at **each** dose level
repeated	at **each** dose level
counted	at **each** dose level
treated	at **each** dose level

{Clinical extraction: from a subject group}

separated	from	**each**	colony
aspirated	from	**each**	mutant
removed	from	**each**	contact

prepared	from	**each**	treated region
withdrawn	from	**each**	sample

3.45 METHODS salient word 8: From

'From' reveals a preoccupation in the Methods sections with the source of data samples, particularly from organisms. *'From'* is involved in embedded passive clauses in complex nominals (a 'reduced-relative' pattern). Most verbs used as reduced relatives have the same essential meaning 'extracted' as in *breast cancer tumours derived from host normal cells*. Similar verbs include: *eluted from, extracted from, harvested from, isolated from, obtained from, prepared from, removed from, taken from*...). We can also see in the following examples similar noun-verb relations to those presented under 'were', where only genetic material tends to be 'extracted':

DNA	was extracted **from**	paired frozen tissue
DNA	was extracted **from**	bone cells using...
Ribonucleic acid	was extracted **from**	PALL cells
mRNA	was extracted **from**	the parent cells
tRNA	was extracted **from**	the exponentially growing cells

One important exception emerges in the reduced relative expression *'obtained from'* which appears to combine both 'extraction from biochemical entity' as well as an empirical 'based on this data source' phraseologies:

{Research data source}

cells	**obtained from**	Dr JH van Dierendonk
data	**obtained from**	the above reaction
cultures	**obtained from**	Sigma Chemical Co.
tissues	**obtained from**	hospital recalls
values	**obtained from**	the previous study

{Clinical extraction}

DNA	**obtained from**	patients
cell lines	**obtained from**	platelet rich plasma

mice	**obtained from**	breeding colonies
tumours	**obtained from**	control mice
A factor	**obtained from**	green tea leaves

'From' in noun phrases generally has the 'extraction' meaning. A notable collocation is '(specific biochemical} cells from {biochemical specific: culture}

trypsinized	cells **from**	monolayer cultures
spleen	cells **from**	tissue culture
tumor	cells **from**	peripheral tissue cultures
mononuclear	cells **from**	control animals
epithelial	cells **from**	immunized mice

3.46 METHODS salient word 10: With.

We have already mentioned the significant role of *'with'* in a collocational framework with *'were'*. Whereas in Titles *'with'* is a salient word used to conjoin similar research processes, in the Methods subcorpus it signals the instrument or medium by which the clinical methodology is achieved. An even more specific phraseology can be found with certain verbs which all have a delimited set of possible instruments:

	{biochemical solution}
were activated with	ethanol
were activated with	an equal amount of saline
were activated with	a cell suspension
were activated with	the culture medium
were activated with	blank human plasma
	{subject-derived serum}
were incubated with	a mouse monoclonal antibody
were incubated with	monoclonal antibodies
were incubated with	antimouse antiserum

were incubated with	test sera
were incubated with	antirat IgG mixture
	{colouring agent}
were stained with	10% ammonium sulphide
were stained with	Alcian blue stain
were stained with	brilliant crystal blue
were stained with	nitro-blue tetrazolium
were stained with	monoclonal antibody

3.5 RESULTS sections

The following results were obtained for grammatical items in Results sections.

Table 15: Results salient grammatical items from the Wordlist program

RANK	**WORD**	**PSC Results Freq**	**% in subcorpus**	**PSC Freq.**	**% in whole corpus**	**Chi sq.**	**Probability=**
16	NO	296	(0.2%)	694	(0.1%)	70.0	0.000
28	IN	3906	(3.3%)	14349	(2.9%)	50.4	0.000
29	DID	176	(0.1%)	395		47.5	0.000
30	NOT	595	(0.5%)	1798	(0.4%)	46.5	0.000
37	HAD	206	(0.2%)	517	(0.1%)	38.2	0.000
41	AFTER	385	(0.3%)	1139	(0.2%)	33.8	0.000
72	THERE	168	(0.1%)	444		25.2	0.000
80	THE	7427	(6.2%)	29122	(5.8%)	23.4	0.000
92	WHEN	184	(0.2%)	518	(0.1%)	20.8	0.000
125	ALL	252	(0.2%)	783	(0.2%)	16.3	0.000

The general phraseology of Results sections is dominated by lexical refocusing, subordination and reporting of quantitative results. We have seen in the discussion of *in*, *did* and *not* above, that Results sections attempt to

evaluate positive and negative results, whereas Abstracts tend to present results (especially negative ones) as quantitative findings.

3.51 RESULTS salient word 1: No.

'No' is the most significant salient word in the Results section, and its role in signalling significant or contradictory data similar to the *'but...'* pattern in Abstracts. *'No'* functions uniquely as a determiner, a usage that is not among the 12 uses of the word in the Cobuild 1995 dictionary. Its most frequent use is in the expression *'there was no significant* {difference / correlation}:

{Empirical statement}	*{Data shape}*		*{Biochemical / clinical}*
There was **no** significant	change	**in**	radiosensitivity
There was **no** significant	difference	**in**	plating efficiency
There was **no** significant	increase	**in**	hydrolysis
There was **no** significant	change	**in**	the time course of efflux
There was **no** significant	variation	**in**	food...consumption

This contrasts with affirmative statements of this kind, which tend to be expressed in the present tense (as discussed above under the item *'there'*). We also find several instances of the passive form of this kind of phrase:

No significant	relationship	was found.
No significant	association	was observed.
No significant	association	was found between tumor grade and LH
No significant	difference	was observed during the time period
No significant	correlation	was observed with respect to rewrite mRNA

The changing preoccupations of the researchers can be seen in the fact that the passive is preferred for research process verbs rather than the clinical verbs observed earlier in the Abstract and Methods sections. When the term 'significant' is not chosen, another evaluative term is necessary with forms of 'to be':

{Empirical evaluation}

There was **no**	**apparent**	effect of diet

There was **no**	**consistent**	pattern across concentration
There was **no**	**detectable**	difference in the incidence of
There was **no**	**strong**	evidence for tumor development

A negative determiner also demonstrates evaluation in relational process verbs:

vaccination	had	**no significant**	effect on the factor
protein inhibitors	had	**no incremental**	effect on tumor growth
ethanol 1%	had	**no apparent**	effect on the p158 cell line
There may	be	**no obvious**	symptoms of cachexia

Other uses of *'no'* reveal the delexical nature of verbs used to report findings. The verb *gave* collocates regularly with the subject *analysis,* while *revealed* corresponds with specific clinical methods:

{analysis}			*{empirical quantification}*
R analysis	**gave**	**no**	indication of allelic losses
SSC P analysis	**gave**	**no**	indication of p52 alterations
analysis of NAK sensitivity	**gave**	**no**	statistical significance correlation

{clinical method}			*{biochemical process}*
screening	**revealed**	**no**	activity
post-mortem examination	**revealed**	**no**	evidence of metastasis
a topographic scan...	**revealed**	**no**	effect within the group

The above patterns could have been expressed using an existential 'there was **no'** (as in the Abstract) but here are used to emphasise the biochemical entity or clinical process initiating the empirical lack of relationship.

3.52 RESULTS salient word 5: Had.

The role of the relational processes *'is a'* and *'have a'* is linked with evaluation in this corpus. *'Had'* is more restricted however, and in the results subcorpus, 'had' serves to signal some degree of quantification rather than qualitative evaluation as for *has / have* in Introductions. The subject often tends to be a biochemical subject:

{Biochemical entity} *{Quantification}*

mice	**had a**	decreased	number of formations
animal tumours	**had a**	greater	mean length
rat liver	**had a**	higher	glucose count
patients	**had a**	lower	frequency
protein	**had a**	more pronounced	effect
infants	**had a**	much lower	susceptibility
controls	**had a**	normal	haryotype enzymes
subjects	**had a**	smaller	body mass

This pattern has also been noted in relation to the determiner *'no'* which can stand in place of the evaluative quantifier, although this expression is limited to biochemical compound subjects with empirical item *'effect'* as head of complement:

the vehicle [=drug]	**had** no	effect	on tumor expression
ZAAf	**had** no	effect	on the reduction of tumor size
treatment of narial cells	**had** no	effect	on weight gain
methanol control	**had** no	effect	on number of implantations
2 weeks experiments	**had** no	effect	on the factor X activator

One fixed collocation emerges in this context: {tumour expression} had significant prognostic value:

Ta-T tumours	<**had** significant prognostic value>
tumor expression	<**had** significant prognostic value>
overexpression of p53	<**had** significant prognostic value>

The inhibitor	<**had** significant prognostic value>
The receptor antagonist ondansetron	<**had** significant prognostic value>

When *'had'* is used as an auxiliary to express the passive perfect, its participle verbs are clinical processes, in direct contrast with the past passive (*'was /were'*) in the Methods section.

electrode	**had** been	allocated
the film	**had** been	deposited
inspection of the electrode	**had** been	electropolymerised
tumour-bearing mice	**had** been	exposed to
rats that	**had** been	treated to.

This is further proof that the past tense can be seen as a marked tense, indicating proximity to current research.

3.53 RESULTS salient word 8: The.

The statistical significance of *'the'* appears to indicate that textual reference to previously mentioned items increases in later stages of the text, a discourse effect that correlates with increased lexical refocusing and rephrasing in later stages of writing. The definite article is obligatory in several collocational framework constructions, and so is a useful indicator of terminological units. Among the more frequent frameworks, we identify the following categories:

Empirical framework:

by **the**	(addition, method, end, presence, production)	**of>**	<(followed, increased, affected, reflected, mediated)

<for **the**	(basis, achievement, accumulation, crossreaction)	**of>**

<in **the**	(presence, size, staging, setting, release, zones, care, levels, absence, range, appearance, relationship)	**of>**

Clinical framework:

<after **the**	(infusion, administration, end, injection, delivery,	**of>**

	implantation, removal)	

Research framework:

<during **the**	(interval, period, intervals, periods)	**of**	(study, observation)>

Measurement framework:

<(consistency, fraction, precision, on the basis, time course, grading)	**of the**	(product, mean, estimation, loss, incidence, 21%, accumulation)	**of the**	(first values, values, body weight, hyperplasmin, dose, cell populations)>

Mixed category (research + empirical + biochemical?)

<(formed, found, calculated, effect)	on **the**	(sensitivity, basis, range)	**of**	(the cell, these results, the data, our data, p-rated hypertosis)>

<in **the**	(absence, presence, care, liver)	**of>**

It can be seen that in all of these frameworks (with the exception of the biochemical sets) all members of the bracketed cluster share some semantic similarity, even though they may not all fall into our rough 5-part category system. This is perhaps not surprising - as Renouf and Sinclair (1991) point out, collocational frameworks depend on their lexical elements to motivate the structure. The regularity with which some are composed confirms the view that prepositions are particularly important to the phraseological specificity of the corpus. The same can also be said of items which have a wide set of uses in one grammatical role but appear to have a unique phraseology as prepositions (such as *to*).

3.54 RESULTS salient word 9: when.

Some forms of subordination (especially signalled by a conjunctive binder) increase in later stages of the research article. *'When'* is used to introduce subordinate clauses detailing a clinical process after a description of research findings. The Results section can be seen to reformulate and re-word clinical experiments already described in the Methods section. The prevalent

structure involves a research process usually expressed by the passive of two verbs *observed* and *obtained*:

{Empirical item}	*{Research process}*	*{Clinical process}*
loss of the film band	was observed **when**	films were photolysed
distinct redistribution	was observed **when**	cells were treated
The results	were obtained **when**	tumors were exposed
Almost identical values	were obtained **when**	(X) was substituted
A greater than 95% yield	was obtained **when**	the equivalent was treated

In Methods sections *'after'* is used to introduce nominalisations of a clinical process, and in Results sections such expressions can be seen to be 'unpacked' into clauses. This can be seen in reduced subordinate clauses especially with the verb *'compared'*:

{Empirical measurement}	*{Clinical items}*	
were significantly reduced	**when** compared to	controls
yielded a 7 fold increase	**when** compared to	the controls
showed superior effects		the same dos

		e

resulted in growth delay	**when** compared with	injection of saline
produced a significant effect	**when** compared with	groups receiving no treatment
infusion was delayed	**when** compared with	groups receiving no SCTT

3.55 RESULTS salient word 10: All.

'All' is a salient word in Results sections. It plays a role in the phraseology of generalisation across the totality of data, and also an important role in lexical reformulation. Of the more regular lexical phrases *'in all cases'* precedes a statement of specific results:

In **all** cases	the medium was supplanted
In **all** cases	normal weight was regained
In **all** cases	the interval returned to baseline
In **all** cases	the relationship ... fell short
In **all** cases	nuclei had upfield shifts

'All other' serves in particular to rephrase items more generally within a taxonomy:

All other	dose groups of males were euthanized
All other	gross observations were checked
All other	microscopic findings were incidental
All other	microvessels showed no change
All other	regions remained the same in sensibility

3.6 DISCUSSION sections

Table 16: Discussion salient grammatical items from the Wordlist program

RANK	WORD	PSC Discussion Freq	% in subcorpus	PSC Freq	% in whole corpus	Chi sq.	Probability=
1	THAT	1381	(1.2%)	3357	(0.7%)	341.8	0.000
2	BE	788	(0.7%)	1825	(0.4%)	225.6	0.000
3	MAY	383	(0.3%)	658	(0.1%)	223.2	0.000
4	IS	1167	(1.0%)	3169	(0.6%)	193.1	0.000
7	OUR	222	(0.2%)	381		129.0	0.000
9	IN	3991	(3.5%)	14349	(2.9%)	116.0	0.000
11	NOT	662	(0.6%)	1798	(0.4%)	108.9	0.000
12	THIS	704	(0.6%)	1997	(0.4%)	96.2	0.000
13	WE	395	(0.3%)	972	(0.2%)	92.9	0.000
14	HAVE	442	(0.4%)	1127	(0.2%)	92.1	0.000

Whereas the phraseology of the Results section is determined largely by refocusing and evaluation of data, the Discussion section can be characterised by considerable lexical reformulation, explanation (by relational processes and explicit signaling), modality and grammatical projection (most often in terms of reporting or referring to previous research).

3.61 DISCUSSION salient word 2: Be.

The high statistical significance of the infinitive *be* is largely due to the presence of large numbers of modal verbs in Discussion sections. We have seen in the discussion above of *'that'* that modality in the evaluation of findings is a very salient feature of Discussion sections. Evaluation takes two distinct forms: external evaluation (commenting on the value of findings for future research) and internal evaluation (commenting on the significance of findings for the present argument). When *'be'* is introduced by *'can'* the

expression tends to be negative, and is uniquely used to express inclusion or exclusion in respect to the 'internal' research model:

analysis	**cannot be**	excluded
range of interactants	**cannot be**	completely excluded
ratio	**cannot be**	ruled out

'Could' tends to indicate either the researchers' ability to evaluate or explain a biochemical fact in terms of 'external' benefits:

{Biochemical process}		*{Empirical explanation / evaluation}*
chemotherapy	**could be**	a potential benefit
chromatography	**could be**	a promising candidate for
tumor expression	**could be**	an appropriate target
This [inhibitor]	**could be**	explained by two steps
This [overexpression]	**could be**	explained as cellular

This variety contrasts markedly with *'must'* which is limited to the collocation *must be due to* (and thus forms an 'internal explanation')

Biochemical / empirical process:		*Biochemical explanation:*
These results	**must be**	due to administration with
These results	**must be**	due to reabsorption
This suggestion	**must be**	due to enzymatic activity
The dispersion	**must be**	due to seasonal variation
This variation	**must be**	due to increased solvoyosis

This rhetorical certainty clearly differs from its exhortative or empathetic uses in the general language ('you must be tired': a significant use in the Cobuild dictionary). In contrast, the modal *should* does tend to be used to persuade or recommend - a similar usage in the general language. Its main difference with other modals in Discussion sections is that the recommended actions tend to be passivised research processes (its uses are generally external: as in: X *should undergo further investigation*):

{Research process}

should be	evaluated
should be	investigated
should be	mentioned
should be	justified

Furthermore, the expression *'it should be noted that'* is used to introduce a finding from current or previous research ('internal' argumentation):

It should be noted that	tumor cell lines are heterogeneous
It should be noted that	others have found higher expression
It should be noted that	...tests have some degree of interdependence
It should be noted that	the degrees of inhibition... did not exceed 70%
It should be noted that	the decay does not take place in a concerted electron transfer

'Would' tends on the other hand to be used in more instances of hypothetical subjectivity than other modals (mostly 'internal' argumentation):

the most likely source	would **be**	expected to return its reactivity
it	would not **be**	wise to allow plasma
stretching modes	would **be**	sufficient
this localisation	would **be**	in agreement with
such a ...mechanism	would **be**	interesting to know

'Will' also introduces evaluation rather than explanation, and emphasises future research (a clear 'external' phraseology):

cytometric analysis	will **be**	required	for different outcomes
samples	will **be**	required	to determine whether
this cohort	will **be**	suitable	
modulation of their kinase level	will **be**	important for...	
tests	will **be**	of limited value	

If these modals are related to their historical 'tensed' categories, it can be seen that there is no correspondence between 'present tense' modals (*can, will, may* - from our discussion below-) and 'past tense' modals (*could, would, must, should*). With the possible exception of *would*, most modals are however used consistently with argument-internal or argument-external verbs.

A even more explicit distinction between evaluative and non-evaluative empirical processes emerges in examples of phase-modality, where the second verb is introduced not as a subordinate clause but as an infinitive 'tensed' by the initial finite. The most frequent is *'appear to be'* (x39 occurrences), which is accompanied by clear examples of comparative evaluation:

This response	**appears to be**	definitely ruled out
These	**appear to be**	significant relationships
These tissues	**appear to be**	very suitable for sequential measurement
This immunoprocess	**appears to be**	much more resistant to cytotoxicity
This detection method	**appears to be**	important in immortalisation

Other expressions share this pattern, such as *'likely to be*' and *'found to be'*:

(biochemical process X}	was found to **be**	considerably more potent
(biochemical process X}	was found to **be**	more reliable
(biochemical process X}	was found to **be**	the best strategy
(biochemical process X}	was found to **be**	much higher

The evaluative pattern is in contrast with that associated with the phase-modal *'need to be'*, which requires a research process as main verb:

Research process *Research process:*

This hypothesis	needs to **be**	formally tested
the new findings	need to **be**	classified
Many more samples	needs to **be**	examined in order to establish
More.. cell tumors	needs to **be**	studied in order to verify whether

These new strategies...	need to **be**	devised

3.62 DISCUSSION salient word 3: May.

We have seen in previous sections that *'may'* is the preferred modal in subordinate clauses after expressions such as *'it is possible that'* and *'it is likely that'*. In most of these expressions, modality corresponds with explicit markers of evaluation. However, outside subordination the majority of the uses of *'may'* appear to function as true 'hedges' by proposing an explanation and indicating to the discourse community that the researchers know it may not be true in all circumstances. Two of the most frequent examples of this are:

{Empirical result}		*{Biochemical explanation}*
ineffectiveness....	**may be related to**	sensitivity
efficiency of this line	**may be related to**	crosstransformation
the more moderate effect	**may be related to**	cell differentiation
lack of bioavailability	**may be due to**	error prone synthesis
deficiency in ..body weight	**may be due to**	direct effects of replication
Another possibility	**may be due to**	inherent differences in age

3.64 DISCUSSION salient word 5: Our.

The statistical significance of *'our'* in Discussion sections is not surprising given that *'we'* is also a Discussion-salient word (discussed above). Personal pronouns are infrequent in the corpus as a whole, and '*our'* signals a shift from impersonal expression to clear signals of 'ownership' of research in the Discussion section:

Our	results	show that
Our	data	show that
Our	study	shows that
Our	findings	show that
Our	studies	show that

Most references to the researchers tend to involve hedging:

Our study	suggests	that
Our study	suggests	indicates
Our study	suggests	demonstrates

However, if the term *'analysis'* is used, no hedge or complement clause is introduced:

Our analysis	focused on a limited subset
Our analysis	was based on immunohistochemical studies
Our analysis	was based on four methods
Our analysis	was to establish criteria for histology
Our analysis	was to understand embedded tissue

Finally, specifying adverbs such as *'clearly'* are used to emphasise the researchers' certainty when no hedging verb is used:

Our results	clearly	indicate
Our results	clearly	demonstrate
Our results	clearly	show that
Our results	strongly	argue that

3.65 DISCUSSION salient word 8: This.

'This' is an important item in the textual development of research articles. As a pronoun, *this* selects an element from previous discourse as the focus of a developing explanation:

This	suggests that...
This	may explain...
This	might explain...
This	is in agreement...
This	is in contrast to...

This use is more common in Methods and Results sections. In Discussions, *this* is more likely to serve as a determiner, reformulating a previous item or proposition as a more general category (for example, expressing a statistical or biochemical fact as a 'result'):

{Research reformulation as anaphoric utterance}

This	result...
This	finding...
This	observation...
This	model...[ambiguous: this may also be interpreted as a 'structure']
This	hypothesis...

This contrasts with less frequent (but more varied) terminological reformulations:

{Biochemical reformulation by superordinate}

This	region...
This	cell line...
This	group...
This	model [as above, this may also be interpreted as a 'hypothesis']
This	protein...
This	type...
This	compound...

This	activity...

In addition, a series of reformulations correspond to specific collocational frameworks, such as *'This* {empirical result} *in* {biochemical / empirical item}':

This	appearance	**in**	parental cells
This	delay	**in**	PMN appearance
This	difference	**in**	rate constant
This	disparity	**in**	degree of suppression
This	increase	**in**	metabolic rate

In the framework *'This...of'* the pattern involves a superordinate empirical item which constitutes the object of measurement rather than a result (as opposed to the pattern above): *'This* (empirical data set} *of* {biochemical / empirical process/entity}':

This	class	**of**	aromatic compounds
This	dose	**of**	chemical...
This	group	**of**	tumours
This	period	**of**	time
This	range	**of**	concentrations

I have omitted one high frequency item that is very frequently used to reformulate results, but is difficult to classify as either research or empirically oriented on the basis of its intrinsic meaning: *this effect*. We have already seen that *effect* has a complex complement structure, accounting for several complex collocational frameworks in Titles and Abstracts (in particular in collocations with *in* and *of*). The word can be used to label observable and measurable phenomena (such as *this motion, this reaction*) and at the same time can be construed as a researcher's interpretation or modelling of results (*this tendency, this frequency*). The word appears to lie somewhere in between *this hypothesis* (a clear research-orientation) and *this activity* (an empirical observation). By reformulating observations as an *effect* the researchers simultaneously explain results and comment on previous data without proposing a new model:

#1 The increased liver weight was reversible.	#2 **This effect** could be the result of increased intracellular glycogens
#1 Treatment with 8-chloro cAMP drastically reduces R1 levels.	#2 **This effect** is even more pronounced in MCF LOA cells
#1 LUMO gap is correlated with downward shift.	#2 **This effect** is misleading. However, some shifts are involved...
#1 Both approaches resulted in 80% inhibition.	#2 **This effect** on ECM degradation indicates that cell UPA is much more efficient.
#1 EFF cells grew slightly faster in MEM.	#2 **This effect** was independent of oestrogens.

To use Halliday's terminology, the clause introduced by *this effect* is an expansion of a previous formulation. The expression differs with research process re-phrasings such as *'This result'* (the most frequent expression used with *this*). *This result* tends to introduce a new research direction which does expand on the previous result but essentially goes beyond it in a reference to research implications:

#1 DNA sequencing of the melanoma revealed that p53 codons... were wild type.	#2 **This result** eliminates the possibility that mutations are germline...it suggests a mutagenic mechanism.
#1 We observe several large AJ-IX positive mRNAs	#2 **This result** may indicate that AJ-IX is a very distant exon.
#1 90% of the carbonium ion was trapped and	#2 **this result** suggests that inorganic phosphate can compete with water to trap the ion.
#1 The reaction.. produces MeOArc.	#2 **This result** is consistent with the partitioning of a common intermediate.
#1 The study .. produced a 23 response rate	#2 but we have not been able to reproduce **this result.**

It can be seen from both of these items that reformulation is not just a process of lexical selection, but also involves the rest of the clause which accompanies the reformulating item. It seems that the meaning of reformulations such as *'this effect'* and '*this result'* depend on the orientation of the following clause. The semantics of a particular word are therefore thrown into sharp relief by its context of use, but can also be seen to be stable in rhetorical terms– at least in the context of a particular genre.

Language in Performance Series No. 22, Tübingen, Gunter Narr Verlag, 270pp.

V. Phraseology and the Discourse Of Science

The main focus of this book has been to examine the specific context of the cancer research article. In previous sections, I proposed that grammatical items are a useful starting point in the analysis of scientific texts. The collocational behaviour of a selection of grammatical items was set out in the preceding chapter in order to relate patterns of phraseology to the style and rhetorical function of the different sections of the research article. I now summarise the main findings of this study and examine some of the implications and limitations of the analysis carried out in this book.

1. Collocations and the Theory of Phraseology

Collocations are words which tend to co-occur in recurrent, recognisable expressions. Our data analysis above shows different collocations are attracted to grammatical items in different types of text. At a basic level of text analysis therefore, I hope to have shown that the comparison of word lists and collocational patterns provides a systematic method of contrasting a specific genre with a general corpus of texts. Collocational patterns thus appear to be fundamental units in the stylistic description of texts.

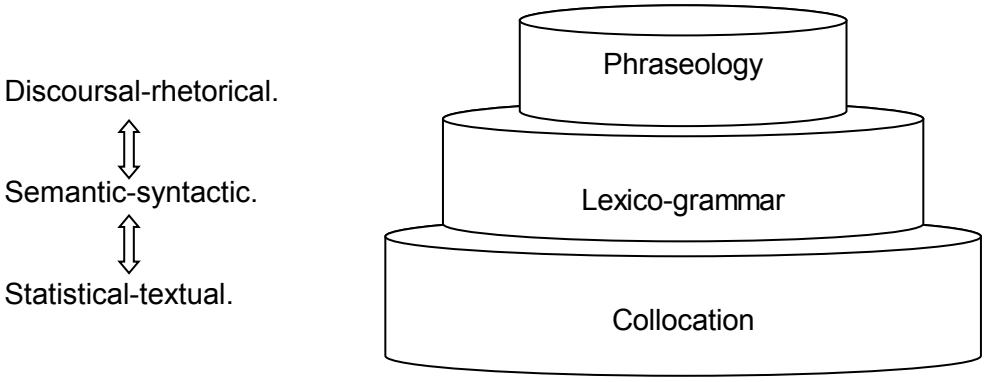
I also hope to have established the notion of collocation within a general theory of language. In phraseology studies, it is generally accepted that clusters of more than one word can reflect a single choice. We have seen in the data analysis above that fixed expressions are often made up of sequences of grammatical items alone, or in combination with high frequency lexical words. In addition, when different lexical items are involved in collocation, the differences of phraseology they exhibit suggest that they are chosen with their role in the larger text in mind. Thus words are chosen not simply for the information they bring along but also for their long-range ability to signal textual relations. These observations appear to confirm the role of grammatical collocation in discourse, and serve to redefine the relationship between the word and the text.

The starting point of my analysis has been to establish a basic 'statistical / textual' definition of collocation. This view of collocation does not pre-define the unit of analysis as a grammatical phrase, but seeks simply to find

significant recurrent expressions. The term 'statistical' derives from Berry-Rogghe's (1970) analysis of statistical collocation and also refers to Sinclair's procedure of relating different distributions of collocation to lexical or grammatical categories (Sinclair 1991). The term 'textual' is used here to suggest that collocations must at first be defined in terms of their textual occurrence, that is to say their use in authentic, naturally occurring texts. However, the analysis I set out above demonstrates that there is more to collocation than word frequency and co-occurrence. We have seen that there are considerable restrictions on expression in science writing, and that semantic sets of low frequency words (lexical clusters) tend to be organised very consistently in specific grammatical patterns, a restriction that is compatible with the 'semantic / syntactic' view of collocation set out by lexicologists such as Howarth (1998) and the systemic grammarians, in particular Hunston and Francis (1998). We have also observed that on many occasions, collocations and lexical phrases are used as specific communicative acts. This corresponds to a 'discoursal / rhetorical' view of fixed expressions, as seen in the work of Nattinger and DeCarrico (1992) and Fernando (1996). Thus collocation is a fundamental notion within a much broader and more complex system of phraseology. I have already noted that this use of the term does not correspond to that used by many lexicologists. Instead this view of phraseology is compatible with the work of Gläser (1998) and Moon (1998a and 1998b). The statistical analysis of collocation is therefore the building block upon which more sophisticated degrees of description and explanation can be based.

Phraseology is the 'preferred way of saying things within a particular discourse'. The notion of phraseology implies much more than inventories of idioms and systems of lexical patterns. Phraseology is a dimension of language use in which patterns of wording (lexico-grammatical patterns) encode semantic views of the world, and at a higher level idioms and lexical phrases have rhetorical and textual roles within a specific discourse. Phraseology is at once a pragmatic dimension of linguistic analysis, and a system of organisation which encompasses more local lexical relationships, namely collocation and the lexico-grammar. I claim that the phraseological analysis of a text should not only involve the identification of specific collocations and idioms, but must also take account of the correspondence between the expression and the discourse within which it has been produced. A visualisation may help to conceptualise the relationship between these three different levels of lexical organisation:

Levels of organisation. Systems of organisation.



The flow chart on the left represents increasingly sophisticated levels of textual description. While these are mutually dependent and inclusive (with collocation providing the basis of all observations at a phraseological level, for example), they correspond to systems of explanation which differ in essential ways (i.e. syntagmatic, semantic and pragmatic systems). By breaking phraseology down into sub-systems and attempting to fix the relationship between such terms as phraseology and lexico-grammar in this way, I am proposing a framework within which it is possible to discuss various levels of lexical expression in a particular text. At the same time, the model distinguishes usefully between descriptive systems, which are often felt to be interdependent, and their corresponding explanatory systems which differ in qualitative terms. I use the terms of this model to summarise my general findings below.

2. Phraseology and Scientific Style.

The analysis of grammatical items in the preceding chapters of this book has revealed a number of interesting properties of the scientific text. From the point of view of genre analysis and *English for Specific Purposes* (ESP), there is much to be said about the role of grammatical collocation and scientific style. The data I set out above show how statistically significant grammatical items can be identified using *Wordlist* (Scott 1993). This provides a list of 'salient' words for each section of the research article (these are summarised in section 4.3 below). Even this relatively simple, mechanical step reveals that the distribution of grammatical items varies

systematically in different rhetorical sections of the article. More generally, I claim that collocational patterns are central to the analysis of register, genre and style. This textual view of collocation is compatible with more recent work on the theoretical framework of lexicogrammar (Halliday 1985) and the phraseological analysis of texts (Moon 1998a and 1998b).

One implication of the data I have presented here is that there is a shared scientific voice or 'phraseological accent' which leads much technical writing to polarise around a number of stock phrases. Fixed expressions ranging from *drug of choice..., yielded modest increases in..., is stable to the action of...* are pervasive in the corpus, but are also at times unusual formulations which are stylistically marked in comparison with general English. While they appear to be normal from the point of view of the science writer, such particular forms of expression stand in marked contrast to alternative ways of putting words to these ideas, a point that is often lost in large-scale corpus analysis. As Halliday (1998) has recently noted, there is a 'favourite clause type' in scientific English. Complexes of two or more clauses are typically compacted as 'things' (noun phrases) in a simple relational clause, the kind of sentence structure that appears to be widespread in scientific writing. He gives an idealised example (1998:190):

Process	**Relation**	**Process**
1 The driver drove the bus too fast down the hill,	**so**	the brakes failed.
2 The driver's overrapid downhill driving of the bus	**caused**	brake failure.

The wording in 2) is an example of Halliday's notion of grammatical metaphor. We have seen in the introduction to this book that grammatical metaphor serves to re-express a complex formulation, taking it generally towards a more nominal mode of expression. In fact, many of the seemingly complex idiomatic expressions we find in the corpus share this underlying property. Thus a *drug of choice* is a behavioural process encoded as a nominal entity, *stable to the action of* is a relational process encoded as an adjectival quality, and *yielded increases in* is an empirical observation of circumstance encoded as a material verb. Halliday claims that such highly distilled structures share the single underlying mechanism of grammatical metaphor (1998: 211). He further points out that far from merely providing novel ways of saying the same thing, grammatical metaphor plays a useful role in the distribution of thematic roles within the clause and at the same

time is a key mechanism in the construction of new meanings. Nominalisation has been noted on many occasions before in science writing, but Halliday has shown that the process is present in a whole series of grammatical expressions and clause types. Other Hallidayan linguists, including Banks (1994) and Derewianka (1994), have pointed out that this shift of expression underpins processes such as modality, hedging and the use of the passive in science writing. Thus from the point of view of phraseology, the underlying tendency to use grammatical metaphor explains to some extent why scientific language appears to be so constrained and so stylistically marked in relation to the general language.

Collocational patterns emerge as a consistent but largely subliminal feature of language. They are specific to the genre and even to the subgenre or section of the text. And those collocations which emerge in our corpus appear for the most part to be consistent with the general stylistic shift of scientific English towards grammatical metaphor. The regularity and widespread nature of much of the phraseology we have observed above is compelling evidence not only for the existence of a discourse community, but for the pervasive influence of community norms on general style and expression. Such consistencies have been identified widely in the literature on genre analysis, and range from the macro-level of the text to small-scale grammatical patterns of usage. Thus Swales' (1990) conception of discourse community relies on large-scale regularities in rhetorical structure, while Myers (1991) examines the consistent use of long-range cohesive devices within the research article genre. On the other hand, Master (1987) examines the role of generic *the* in research articles, and Salager-Meyer (1992) and others examine lexical metaphor, the rhetorical role of tense and verb form in science texts and other micro- textual features. I suggest that the collocational patterns we have seen above (including the use of idioms, fixed expressions and other formulae) provide a useful intermediary stage of analysis between the macro and the micro levels of linguistic description. Collocation is the link between the word on the one hand and the text on another.

Collocations appear to confirm the existence of a discourse community. Their very consistent nature suggests that collocations have a central role to play in discourse, at a metaphorical level in terms of reformulating ideas but also, to use Halliday's terms, at the level of textual organisation and interpersonal expression. Nevertheless, this picture is complicated by the fact that the research article genre does not have a single monolithic style, or lexico-grammar, with entirely predictable features. The sheer variety of graphic presentation from one research specialism to another is a useful reminder of the complexity and heterogeneous nature of scientific discourse. The regularity and pervasive nature of collocation appears to be incompatible

with the intuition that an individual's use of language is inherently unique and creative. While presentation and format are matters of conscious editorial control within different research articles, collocational style is presumably not a conscious product of composition or of editing. Instead, it is likely that the collocational coherence of a text is an acquired characteristic derived by the writer from wide reading and sub-conscious attempts to conform to the norm of speech in the scientific community.

I have pointed out above that the Pharmaceutical Sciences Corpus includes a wide variety of different specialisms even within the specific field of cancer research. Even texts within the same journal cover very different areas of research, and the authors originate from different institutions and language backgrounds. So it must be the case that examples of collocational regularity across these widely different research specialisms (and across a broad range of periodicals) represent a form of coherent scientific style. The term I propose for these expressions is *generic collocation.* Thus just as the discourse community has its system of genres and technical jargon, it may also develop a more subtle set of identifying expressions, at least in its formal modes of written communication. It does not appear enough however to suggest that collocations and phraseology are dependent on style and interpersonal factors such as similar rhetorical functions. I have suggested above that phraseology may have an important role to play in the textual development of meaning, and so any explanation of the consistent style must in some respects return to the preoccupation of terminologists and attempt to relate the 'preferred way of saying things' with the prevalent knowledge structure of science.

More recently, Lemke (1998) has shown that several genres are present within a single text, and that it would be an oversimplification to see scientific style as purely limited to a specific genre within the broader language system. Despite the collocational specificity of many of the expressions we have examined above, there is no reason to believe that scientific texts are wholly separate from the general language or that they do not interact with or derive new modes of expression from everyday speech. Indeed, Halliday and Martin (1993) have consistently argued that the general language is itself imbued with the phraseology of several competing technical registers, from the language of science and religion to that of business and journalism:

> Every text, from the discourses of technocracy and bureaucracy to the television magazine and the blurb on the back of the cereal packet, is in some way affected by the modes of meaning that evolved as the scaffolding for scientific knowledge... In other words, the language of science has become the language of literacy (Halliday and Martin 1993:11)

Halliday and Martin see the influence of scientific discourse as pervasive in society, especially in the context of advanced and higher education. Their thesis has been to alert educational authorities to this influence so that students from non-literate backgrounds can deal with technical language. While other forms of discourse may be equally as influential (such as the discourse of commerce), scientific discourse can be seen to operate in a large number of genres that are ultimately derived from research articles. As we saw in the PSC survey of scientists in chapter 2, research articles compete for the reader's attention with review articles, experimental articles, accelerated communications, 'popular' science articles (in *Nature* etc.) and indexing abstracts. But one can also note the important role of the 'grey literature' (Auger 1979); that is, of grant proposals and the reports of the research funding councils, and the press releases of the major cancer charities.

Specialist research articles have adapted very specialised ways of processing scientific knowledge. But science as a human activity is embodied in discourse, not just in research articles and the discourse of science is appropriated by various groups rather than produced or reproduced in texts.

3. The Lexico-grammar of the Scientific Research Article.

The theory of lexicogrammar is based on the observation that different words tend to have unique grammatical relations, and that extended expressions tend to include only those items which have the same semantic properties. This book has attempted to construct the essential elements of a lexico-grammar of the research article genre, at least in the field of cancer research.

To present a summary of the lexico-grammar of research articles here would belie the complexity of the data. Nevertheless, there are some general correspondences between grammatical items on the one hand and the communicative functions of each section in the corpus. The picture of a homogenous grammar extending from the Title to the Discussions section fades away, and we are left with highly specific grammatical subsystems for each of the rhetorical sections of the article. These remarks become even more significant, when one considers that most of the 'science' in the research article is reformulated from one section to the next, and that the text is in effect a cyclical series of more or less complex paraphrases and re-evaluations of the same data. The differences in wording between different sections must therefore be interpreted in terms of the textual and interpersonal functions of the text rather than simply in terms of propositional information.

Introductions, for example, involve the lion's share of infinitive clauses of projection (clauses introduced by 'to', e.g. *has been shown t*o... + non-finite verb), while projection in Abstracts and Discussion sections is typically finite (*it has been shown that* + finite verb). In addition, we have seen above that even the same salient items in different rhetorical sections have subtle but consistent variations in use. For example, while there is significant negative polarity in both Abstracts and Results sections (expressed by *did not*), Abstracts summarise the quantity of negative results (*did not decrease significantly*), while Results sections compare data and explain negative results in terms of quality (*did not result in significant metastasis*). Generally speaking, grammatical items in cancer research articles tend to have a much more restricted set of uses than in the general language (at least in comparison with items listed in the Cobuild dictionary). Thus despite differences between conventional sections, some individual grammatical items share associated phraseological roles throughout the corpus. This involves such features as the construction of nominal groups (where 'of' is a significant item), signalling of negative results ('but'), the reformulation of immediately neighbouring discourse ('this'), evaluation in relational clauses (following 'is, have'), research- or empirically oriented clause complexes ('that' or 'to'), passives ('been'), the quantification of clinical processes ('at'), the qualification of effects or results ('in'), the expression of modality and hedging ('be') and indirect impersonal metaphor ('it'). Thus while a grammatical item in the general language may have a largely unpredictable set of contexts, the corpus allows us to infer a very specific phraseology and system of lexico-grammatical relations for these words.

However, the lexical and semantic structure of the research article becomes much more predictable when we examine coherent subsections of the corpus. For example, the typical phraseology of Titles centres on prepositions such as of which are used to form complex nominal groups. The focus of research in Titles tends to be to the left of the expression with an empirical or biochemical finding in thematic position with post-modifying phrases tending to express clinical methodology. If the left-hand item is a semi-technical noun, such as evaluation, relation, effects then this item serves as the methodological focus of research rather than a biochemical entity, although this entity or process must then be expressed as the next element (i.e. is not head of the noun group). While this is the dominant phrase structure, a minority of Titles also involve active clauses, which usually involve an attributive clause, serving as an immediate evaluation of results:

Titles

inhibition effects **of** chemotherapy **on** metastases (complex biochemical nominal)
Evaluation **of** prognostic factors **in** breast cancer (complex research nominal)
tobacco as a risk factor **for** lung cancer (nominal with goal)
The relation between clinical **and** histological outcome... (framework with conjunction)
pS2 is an independent factor **of** good prognosis **in** primary breast cancer (evaluation)

In contrast, salient expressions in Abstracts represent grammatical compaction (relative clauses and hypotactic expansions which define the scope of reference of a the main nominal expression) and the quantitative reporting of data shapes (rising, falling, stable or negative statistical results) together with other past-tense findings:

Abstracts

the mechanism **of** action **of** {compound Y} **was** shown **to** {+ empirical process} (complex nominal expression of findings)
there was a significant increase **in** toxicity (quantitative report)
It is concluded **that** propagation **did** not increase (impersonal expression of quantitative report)
subjects **who** receive active management (fixed embedded clause)
both normal and tumor cells (framework with co-ordinate conjunction).

Introductions in turn contain perhaps the longest stretches of consistent phraseology, generally reformulating previous research or evaluating established concepts (in the present and present perfect) or announcing action-oriented events (research aims and intended methodology expressed in the past tense). Such events tend to be associated with *to*- and *that*-clause projections:

Introductions

p53 gene resistance **has been** reported (fixed expression of report)
PIMO **has** received little attention (fixed expression of report)
studies **have** shown **that**... (fixed expression of report)
is an effective inhibitor (expression of evaluation)
(Compound X) **is** stable to the action **of** (Compound Y) (expression of empirical result)
use of agents **such** as dismutase (refocusing previous item)
it was also found that (reporting previous research)

In this study **we** examine (fixed expression of report)
the purpose **of** the present study was **to** expand data (fixed expression).

Methods sections contain a variety of fixed expressions, and their phraseology is principally concerned with the circumstances of clinical procedure such as sequences, rates of change and clinical extractions from one data source to another. The past passive also becomes prevalent in the reporting of (recent) clinical events in this section:

Methods

aminids **were** censored **from** the organs (idiosyncratic expression of procedure)
was examined **for** external defects (clinical expression)
at each dose level (procedure)
(Compound Y) **was then** added dropwise (clinical expression)
was collected **and** concentrated (clinical sequence)
(data set) calculated **from** the bootstrap samples 24h **after** exposure to (fixed expression of procedure)

The salient expressions of Results sections are predominantly concerned with qualitative reporting, reformulation and comparison of positive and negative data. Prepositions such as *in* which are used to introduce clinical data sets elsewhere (for example in Abstracts and Titles) are now used in nominal modifiers expressing empirical observations. Grammatical projections (in *that* and *to*) are replaced by existential impersonal expressions of report (using *there is, there are*) or expansion clauses (introduced by *when*):

Results

There was no significant change **in** radiosensitivity (qualitative report)
controls **did not** show RT activity (qualitative report)
mice **had** a decreased number of formations (quantitative report)
it appears that **there** are considerable differences (qualitative report)
after **the** infusion **of** (clinical framework)
no activity was observed **when** (X) was incubated (qualitative research report of clinical process).

Finally, Discussion sections typically express overt evaluation (referring to *we* and the use of projections with *is*) and explanation of data reformulated as empirical rather than biochemical processes (notably after *in*). As might be

expected in research papers, the Discussion section refocuses attention on a conceptual research model and reformulates empirical observations as cognitive / research-oriented nouns: *models, hypotheses* and *strategies*. Clause projections in *that* becomes prevalent (*that* introduces cognitive research processes as opposed to *to* which tends to introduce biochemical events) and modal verbs are used in widespread hedging:

Discussion

data suggests **that** reactive oxygen would **be** important (modified report of results)
This result **may be** related to bleeding tendency (modified explanation)
It **is** interesting to note **that** (modified research report)
increasing data does **not** result **in** any further enhancement (qualitative report)
This evidence suggests **that** (including reformulation)
we have found **that** (report)

Although I have used the words 'typical' or 'prototypical' in reference to these expressions, it is perhaps more accurate to describe this as outstanding phraseology. I chose the term 'salient' to capture the idea that these expressions are only typical of those elements of style which are in some way deviant from the rest of the corpus. This is because the *Wordlist* comparison emphasises extreme differences in the corpus, and although concordance analysis does suggest some similarities, it sheds little light on phrases which may be used consistently from one section to the next. The expressions listed above are in fact untypical, at least in respect to the corpus as a whole, although they are of course prototypical of the section of the text which they represent. It has to be noted therefore that a degree of potential consistency may have been overlooked by the large-scale statistical analysis of differences in the corpus.

Although grammatical collocations are useful for identifying longer stretches of phraseology, it has not yet been proven that they represent the overriding phraseology of the text as a whole. The listing I present above represents an extreme generalisation and it is difficult to gauge from this the proportion of any one individual text which may be made up of prototypical or outstanding phraseology. In particular, it is important to relate these findings above to individual texts. To examine this dimension of text analysis, I have annotated below a Discussion section from one article: "*Bioreversible Protection for the Phospho Group*", a paper donated to the corpus by the lead-author S. Freeman and originally published in the *Journal of the Chemical Society* (Vol.13, 1991). A rough indication of the extent to which such a text conforms to the typical lexico-grammar of the corpus can

be shown by graphically identifying those items mentioned as salient in the PSC in bold, and at the same time indicating lexical items which are usually collocations of salient items in the corpus (underlined). (Triangle brackets are used to separate phrases found in the general phraseology from those which appear to be untypical. Thus bold items outside triangle brackets indicate non-typical uses of grammatical items identified in the corpus):

Comparison of typical PSC phraseology with a pharmaceutical Discussion section.

<The ready removal **of** the 4-acetoxybenzyl groups> with carboxyesterase <suggests **that the** 4-acyloxybenzyl diesters **may be** useful bioreversible derivatives **of** the phospho group>. <**The** lower reactivity **of the** monoester> with carboxyesterase <**when** compared with **the** diester>, <could **be** exploited> **to** provide <a sustained release **of** parent drug>. **In** theory, once inside **the** cell, **the** lipophilic diester would readily <yield **the** anionic monoester>, which being charged <would **be** trapped> **and** hence serve as <a reservoir **for the** parent drug>. <**This** bioreversible protecting group could also **have** applications **in** synthesis>, with **the** phospho moiety being liberated under very mild conditions avoiding <**the** common methods **of** high pressure hydrogenation>,3 strong acid14 or trimethylsilylbromide.15
Although the products <derived **from** the phospho group **of** the diester (1) are known>, the fate **of** the benzyl group <**is** more complex> with only ~< 30% **of** the product derived **from the** proposed carbonium ion> being present as 4-hydroxybenzyl alcohol <**at** early time points>. Instead **of** reacting with water, <the carbonium ion **may be** trapped by another nucleophile>, **and** possibilities include **the** enzyme, products or buffer. <**The** reaction profile **for the** decomposition **of** triester (1) with carboxyesterase **is** very similar **to that of** monoester (2)> (Figure 1). **For** (1), <two equivalents **of the** carbonium ion> are generated, which <does **not** lower catalytic efficiency>, <**this** suggesting **that this** intermediate does **not** react with enzyme>. <**In** a related reaction[16] the benzyl carbonium ion generated **from the** solvolysis **of** diphenyl benzyl phosphate **in** phenol> is trapped by electrophilic aromatic substitution <to give 2- and 4-benzylphenol>. <An analogous reaction **of the** 4-hydroxybenzylcarbonium ion> with 4-hydroxybenzyl alcohol would give 3-(4'-hydroxybenzyl)-4-hydroxybenzyl alcohol, however **the** 1H n.m.r. spectrum only suggested 1,4- disubstituted products. To investigate <**the** involvement **of the** buffer> <**the** reaction **of** (1)> with <5 units **of** carboxyesterase> <**was** repeated using 0.01 M phosphate buffer>. <**At** all time points more than 90% **of** the carbonium ion **was** trapped as 4-hydroxybenzyl alcohol> **and** <**this** result suggests **that**> with the original 0.1 M buffer, <inorganic phosphate **can** compete> with water **to** trap **the** carbonium ion. Although <**we have** yet to prepare a standard>, unassigned peaks <**in the** n.m.r. spectra **of the** reaction mixture> with 0.1M buffer are dP 3.72 ppm **and** dH 7.26 (2H, d, JHH 8.4), 6.81 (2H, d, JHH 8.4) **and** 4.64 (2H, d, JPH 5.4) consistent with

4- hydroxybenzyl phosphate, which <**has** an approximate half life **of** 1 h.> <**The** monoanion **of** benzyl phosphate> <**is** reported **to** hydrolyse> with P-O cleavage with <a half-life **of** 86 h **at** 75.6 oC **and** pH> 7.17,18 <**The** higher reactivity **of** 4-hydroxybenzyl phosphate suggests a change **in** mechanism>, with **the** electron-donating hydroxy group promoting C-O cleavage. Studies are **in** progress **to** optimise <**the** stability **and** bioactivation **of the** 4-acyloxybenzyl phosphodiesters>, <**for** both drug delivery **and** as a synthetic method>, by altering <**the** nature **of the** acyl group>. **The** potential problems associated with <**the** release **of** a highly reactive benzyl carbonium ion <**have been** outlined>,6 <**and** methods **to** trap **this** intermediate> internally are being investigated.

This visual identification of collocations allows us to contrast those features that are typical of cancer research articles in general (the corpus) with features which appear to be distinctive in the style of this particular text. It can be seen that approximately 30% of the text (151 items out of 496) is not involved in the typical phraseology identified in our main corpus analysis. At the same time, this visualisation shows that many collocations run into each other and are interdependent. Any two bold items included in the same brackets appear to share lexical collocations, and presumably also collocate as an extended expression. Such sequences of interlocking items are termed *collocational cascades* (Gledhill 1995a): collocational patterns which extend from a node to a collocate and on again to another node (in other words, chains of shared collocates).

What is of interest in terms of genre analysis is the extent to which this text differs from the corpus-based norm. The Discussion section observed here has features of language which are typical of other sections (such as a high number of projecting clauses). But there are also features which are very untypical, including expansion clauses introduced by *to* (as a synonym of '*in order to'*) in dependent clauses signaling a circumstantial aim or consequence. This feature does not occur prominently in other Discussion sections or in fact any other section in the corpus (Introductions favour *to*-complement clauses or projections, such as *It is important to…*, *The aim was to*...). The text also uses an unexpectedly large number of non-finite clauses after *with* (in an expansion + *ing*). However, the most striking feature of this text is the number of reduced relative clauses: *mild conditions* [[*avoiding the common methods of high pressure hydrogenation*…]], *the phospho moiety* [[*being liberated...*]], *yield the anionic monoester* [[*which being charged...*]]. The final example here involves the pronoun *which,* which happens to be the 17th most salient grammatical item in the Discussions subcorpus (468 uses out of 1422). This suggests that non-restricted relative clauses are also typical of other Discussion sections. This differs from Abstracts, which use explicit (non-reduced) relative pronouns (*who, that*) more often in defining relative

clauses attached to a noun. In other words, Abstracts use restricted relative constructions and tend to reformulate and summarise findings first presented and evaluated elsewhere, usually in Results sections. Discussion sections, on the other hand, prefer to use dependent clauses which add new information, extending the thematic range of the clause as a whole. Reduced relative clauses such as the ones we find here do not appear to be frequent in other Discussion sections however (only five *–ing* verb forms appear in the first 1000 salient items in that subcorpus). Thus reduced dependent *–ing* clauses and dependent circumstantial clauses introduced by *to* (‘*in order to*’) appear to be an idiosyncratic feature of the individual style of this text rather than a feature of the genre as a whole.

One of the more fundamental findings to emerge in our study is that the phraseology in the corpus tends to correspond very consistently to a small set of dominant semantic categories. In the Pharmaceutical Sciences Corpus most lexical items were found to belong to four main process types: RESEARCH, EMPIRICAL, CLINICAL and BIOCHEMICAL. These four dimensions form a continuum in which they represent the relative involvement of the author in the scientific activity (either in experimentation or writing up). RESEARCH processes can be seen as the most overt expressions of an author’s mental or behavioural involvement, and BIOCHEMICAL processes are seen as the most distant from the author (representing a chemical, material process with no overt external agent).

Increasing ‘autonomy’	**Increasing ‘intervention’**
RESEARCH	RESEARCH
↓	↑
EMPIRICAL	EMPIRICAL
↓	↑
CLINICAL	CLINICAL
↓	↑
BIOCHEMICAL	BIOCHEMICAL

As might be expected, these semantic categories correspond indirectly to the fundamental processes identified in Halliday’s (1985) grammar of transitivity (the main processes in the general language are: material, relational, verbal, mental, behavioural, existential). As with Halliday’s terms, our process types are open to reformulation as grammatical metaphors (for example, processes expressed as events etc). Although the terminology does not correspond directly, it can be seen that the process types identified in the corpus can be

realised as entities (prototypically nouns), qualities (prototypically adjectives), events (prototypically verbs) and circumstances (prototypically adverbs and prepositional groups).

Thus semantic categories emerged at all points in the corpus analysis as collocates of grammatical items and longer stretches of phraseology. Such 'clusters' are a well-documented feature of collocation, and are often seen to coincide with small changes in grammatical formulation (Sinclair 1991, Carter 1997). For example, in Methods sections (but not elsewhere) the past passive phraseology <*were* + past participle> involves mostly clinical verbs (*were sliced, incubated, filtered*) or empirical verbs with associated prepositions (*were increased at, identified as, determined with*). Yet the passive in other sections is expressed in the simple or perfective present tense, and is dominated by research process verbs (*is believed to be, are observed, is concluded that*). A simple interrelation between lexical items and grammatical collocations can be seen in the framework <*were _ by X*> which involves only statistical tests: *X* ***were*** *analysed by Student's t-test*, while the framework <*were _ with Y*> involves only instruments of methods: *Y* ***were*** *determined with NMR spectroscopy*. Another example from the PSC involves the interdependence of verb form and phraseology. As we have seen in a discussion of *there is / there was* (in the analysis of the adverb / pronoun '*there*'), statements of given fact about biochemical entities are likely to be in the present tense (indirect observations), while statements involving research and empirical processes are likely to be in the past tense (direct observations). However, some evidence suggests that the phraseology is constrained on a more specific lexical level. For example we saw above that the subject of a past tense phrasal verb *'led to'* is always a research-oriented process (*these observations led to...*) while the subject of the present tense form *'leads to'* is always a biochemical or empirical process (*response to DMT damage leads to...*). Thus, it is also possible that tense correlates with lexical and semantic categories as well as the broader rhetorical generalisations postulated by linguists such as Oster (1981) and Malcolm (1987). The general implication may be that grammatical features which are often seen in terms of open or free choice are in fact determined as obligatory parts of a complex, extended lexical expression, as first posited by Sinclair (1991)..

The principle of a lexico-grammatical system becomes immediately apparent when one examines the middle ground between lexical and grammatical items, including high frequency lexical items and what are known as non-technical words. I have shown elsewhere that non-technical lexical items in science writing are involved in highly specific and consistent grammatical systems. These items are used in a lexical sub-system that may

be independent of the general language. For example, in Gledhill (1997) I examined the lexical phraseology of high frequency nouns and verbs in the corpus. I found that the collocational patterns of verbs such as *show* and *demonstrate* display very consistent grammatical differences. *Show* is typically involved with non-finite projections of the type *X has been shown to* {+ empirical finding}, while *demonstrated* is used with a simple complement or a finite projection *it has been demonstrated that* {+ finite statement of biochemical fact}. But a further unexpected difference involves the polarity of the two verbs: *demonstrated* regularly introduces negative results, either expressed as *failure* (*we have failed to demonstrate X...*) or as a simple negative (*we have demonstrated that X is not effective in the treatment of Y*). The verb is therefore co-selected as part of an extended expression. Putting it another way, the verb *demonstrated* is 'reserved' for the expression of negative results, almost as though the verb is used as part of an extended communicative signal and exists in opposition to more neutral verbs such as *show*.

These instances are complicated by the fact that in a similar corpus of scientific texts in French, the usual translation equivalents of these verbs (*montrer, démontrer*) do not display the same lexico-grammatical properties (Gledhill 1999). The French system involves a verb which has no translation equivalent in English *préciser*, whose use lies somewhere between *indicate* (French *indiquer*) and *show*. The meaning of the verb *demonstrate* in scientific English involves a notion of contrast (not necessarily negative contrast). But there is no such nuance in the French use of the verb *démontrer*. Our understanding of these verbs must therefore depend on our deeper recognition of the underlying phraseological impact of the word as part of an extended expression. While one might expect a general underlying pattern to emerge across different languages within the discourse community of scientists, it appears that French and English science writing may have developed their own specific discourses, with a variety of lexical items employed to express very sophisticated but also very consistent phraseological nuances. If these observations on phraseological patterns do not correspond with the general language, then translation appears to be an more difficult task than is ordinarily assumed, since even non-technical lexical items can be seen to be non-equivalent on a basic phraseological level. Although further work is necessary on inter-cultural and inter-discoursal aspects of collocation, it is clear that these features of the lexico-grammar are systematic but also unpredictable. A collocational pattern is unpredictable in the sense that a native speaker is largely unaware of the consistency of the pattern. However, speakers may be aware of the general phraseological effects of the word, and may associate the phraseological

patterns of the word subconsciously with its connotative meanings. Such a principle is the basis of recent corpus-based dictionary projects, as pointed out by Sinclair (1991).

Generally speaking, linguists such as Hunston and Francis (1998) have found that changes in grammatical sequence tend to involve the formation of coherent, consistent groups of lexical collocates. Such correspondences between global grammatical choice and lexical phraseology are fundamental features of Halliday's notion of lexico-grammar (Halliday 1985). As Francis (1993) puts it:

> 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).

We have seen in chapter 2 that there is a body of linguistic theory that sees such patterns as central to the way discourse is *construed*, or to reformulate Halliday (1985), how we build and interpret the world through discourse. The neo-Firthian view of language set out throughout this book sees the semantics of the word as textually distributed and syntax as intimately linked with lexical knowledge. In the specific context of cancer research articles, knowledge of phraseology involves knowing which tense to use in expressing biochemical and research processes and, to give a very specific example, even a subconscious knowledge of duality in the discipline in the use of basic co-ordinating conjunctions. Phraseological knowledge can be seen as a central factor in the process of writing and reading in this specialist field. In this regard, Francis (1993) has argued that such knowledge is a key mechanism by which we move 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 (Francis 1993:122)

Given this view, that meanings acquire their own wordings, we can therefore conceive of the broader system of phraseology as the set of linguistic forms motivated by rhetorical aims and which further shape the discourse. It follows that the collocational patterns we have identified are formulated in previous text and must have a role in the processing of the text as a whole. The intertextual function of collocation is therefore apparent. Clearly any

changes in phraseology introduced by the author or any deviations from the collocational cascade must have consequences for concepts throughout a running text, as we have demonstrated on several occasions in this book in the analysis of grammatical reformulation.

4. The Role of Grammatical Items in Collocation.

Although grammatical items tend to occupy similar ranks of frequency in a variety of texts and word counts (for example those based on large text corpora such as the *British National Corpus* and the *Bank of English*), this study claims that their use is more predictable in terms of context and function than has been previously suggested. This is because any variations in basic word lists come into sharp focus when the collocational behaviour of these items is considered at a further stage of analysis. It appears from our analysis above that conventional formulations remain consistent within each section of the research article, and that each salient grammatical item tends to contract a different set of collocations from one subsection to the next.

One reason for this is that the communicative goals and semantic concerns of the genre lead to a delimited set of linguistic expressions. When these goals change, the phraseological resources of the text change at the same time. Collocations involving grammatical items are thus consistent indicators of long-range relations between texts. They are usually stable from one text to the next (i.e. within the subcorpus of Abstracts or Introductions etc.), but differ from one section of the article to another. Collocational variation across rhetorical sections affects many areas of grammar and discourse in the corpus, largely because the items that are found to be salient cover a number of grammatical categories. This is not a trivial observation. If the statistical counts are well conceived and accurate, then the rhetorical sections of research articles appear to be very different in terms of a wide variety of grammatical constructions - a point not often realised in those corpus studies which classify the whole text as a single register or text-type (a recent exception has been Biber, Conrad and Reppen 1998).

The lexico-grammatical patterns of research articles show that collocation is not an accidental property but a fundamental characteristic of the genre, as central as such features as rhetorical moves, thematic progression and clause structure. It is interesting to observe that these global features of text tended to dominate the discussion of genre analysis before the advent of computer-based corpus linguistics (for example, Nwogu 1989, Wikberg 1990, Mauranen 1993). It now appears that corpus-based studies have shifted the emphasis of analysis to the micro-level of the genre. It is now possible to

posit generic features of a text with much more certainty than earlier work. There has recently been a considerable amount of research on lexical collocation in technical genres (as in the work of Howarth 1996 and Pearson 1998) or on syndromes of inter-related grammatical categories in the comparison of broader registers (Biber, Conrad and Reppen 1998). Only a small number of studies have begun to examine the distribution of grammatical collocations in a specialised genre, and none have established a comparative analysis of collocation in sub-sections of a text. While the study presented here shares similar methods with many computer-based studies of authorship and information retrieval (for example Ager et al. 1979, Moskovitch and Caplan 1979, Harris 1985, Phillips 1989, Ahmad et al. 1991 and Ide 1993), few of these have focused on grammatical collocation as a means of 'trawling' or fishing out the phraseological properties of the text. The aim of my analysis is therefore to balance those studies of genre which concentrate on the macro-structure of texts (especially within ESP), and also to provide an alternative contribution to mainstream work on the language of science, which has tended to see collocations as an extension of terminology rather than as a feature of text.

Recent studies of corpora of the general language (Sinclair 1991) have begun to challenge the traditional way of seeing grammatical items. Whereas lexical items vary in frequency and distribution across a variety of topics and genres, high frequency grammatical items are assumed to remain the same. Yet much of the evidence I have presented in this book suggests that this picture is misleading. The interaction between a grammatical item and a cluster of semantically-related lexical items suggests that grammatical words should be seen not only as closed-class or high-frequency items, but also as the fundamental elements of organisation in phraseological units. Many grammatical items do of course lack propositional meaning when considered in isolation, but it is important to consider the role of grammatical words within longer phrases and their function in the grammatical reformulation of the text. I have suggested above that grammatical items provide an efficient way of arriving at a description of the most typical phraseology of the genre. And we have also seen that grammatical items and grammatical reformulation have an important role to play in Halliday's theory of grammatical metaphor, that is to say in the formation of textual meaning. When considered from this perspective, it becomes clear that grammatical items and their attendant phraseology have an important role to play in the textual and interpersonal functions of the text.

We have seen that grammatical items are present in the most fundamental phraseology of the Pharmaceutical Sciences Corpus, including such basic expressions as *we conclude that...*, *[compound X] has been shown to*

[dimerize, express, flip...]..., these findings demonstrate that.... These correspond to Nattinger and DeCarrico's (1992) notion of the lexical phrase. Rather than expressing propositional information through terminology, these expressions represent the fundamental style of the text and have specific rhetorical functions. Their textual roles range from reformulating as grammatical metaphors, signalling modality, forming hedged and modal phrases, and refocusing previous discourse. Such expressions are not often seen as prototypical examples of science writing. However, the corpus evidence suggests that grammatical items within lexical phrases are the most stable features of language in the research article. This is partly a consequence of the processes of grammatical metaphor I cited above, but it can also be seen that many of these expressions have very specific phraseological properties which differ markedly from their general-language equivalents.

I have concentrated throughout this book on grammatical collocation (grammatical items collocating with lexical clusters), collocational frameworks (collocations involving more than one grammatical item) and colligation (collocation between grammatical categories). These forms can be contrasted with lexical collocation, for example nominals such as *total synthesis* and *active physiological management*. Lexical collocation is an important feature of scientific terminology. However, lexical collocations do not appear to have the same range or distribution of use as those expressions which involve a grammatical item. As we have argued above, grammatical words play an important role in reformulation and re-wording. Halliday identifies several instances of grammatical metaphors, and all happen to involve grammatical items: *the movement of planets, the instability of diamond, resulted in brake failure, leads to X..., the fact of Y...* (1998: 309-210). It appears that many features of grammatical metaphor involve prepositions, and prepositions have caught the attention of linguists in previous studies (Sastri 1968). This general form of reformulation accounts for the high frequency of prepositions in the PSC word list when compared with the general language (c.f. Appendix 1). We have seen similar instances in a number of areas in the corpus, in particular in impersonal projecting clauses (with conjunctive *that* and *to*) and the passive (involving forms of the verb *to be*). In addition, the mechanisms of 'alternation' in science texts were identified as important processes by Pettinari (1982). These processes correspond to Sager et al.'s (1980) observation that while certain terms can involve basic grammatical reformulation (*drug pusher / a pusher of drugs, measles vaccine / a vaccine for measles*), other more established terms appear to be grammatically fixed (*jet engine / ?the engine of a jet, long-term memory / ?memory for the long term*). This is also reflected in Fischer's

(1998) discussion of neologism and lexical change in the general language, in which the range of successful nominal compounds which involve lexical modifiers (*mind-bending complexity, grant-maintained school, wide-bodied jet*) tends to be greater than compounds involving complex grammatical relations (*just-in-time, hands-on, us-versus-them*). Grammatical collocations thus seem to be central to style and reformulation in the text, while lexical collocations (especially nominal compounds) are represent a system of more-or-less frozen established terms.

In her analysis of the reformulation of idiomatic expressions, Moon (1996) finds that of all the items used in common expressions, grammatical items tend to be the most fixed. This is a departure from the traditional lexicological view of a phrase or fixed expression, in which lexical words are seen as the most useful entries for classification in dictionaries. Conversely, many of the examples in the previous chapter show that while the number of lexical items in a cluster is variable, the grammatical items in a collocational framework are integral parts of the expression. As I noted in chapter two, it is clear that grammatical items and high frequency 'non-technical' words are clues for decoding the scientific research article, and may provide a significant feature of recognition for expert readers. In a study on the readability of scientific texts Clarke and Nation (1980) point out that for non-expert readers, grammatical and high frequency lexical items are the only items they are able to recognise, and their understanding of the text will depend on a coherent reading of collocational patterns in what is essentially an approximation of a cloze-test.

Yet this view of high frequency items has not often been recognised, as I argued in chapter 3. Even Halliday and Hasan (1976) claimed that high frequency lexical items such as *go, man, know* or *way* 'can hardly be said to contract significant cohesive relations, because they go with anything at all.' (1976:290). They also claimed that 'the higher the frequency of a lexical item... the smaller the part it plays in lexical cohesion in texts' (1976:290). Many linguists appear to similarly believe that higher frequency words (grammatical items) are of little interest in the meaning creation of the text, and most large scale analyses of corpora tend to eliminate grammatical items by imposing 'stop-lists'. Yet I hope to have demonstrated that grammatical items play a important role in a number of discourse features of the text (especially in the guise of lexical phrases). Although admittedly Halliday and Hasan were talking about long-range features of textuality, I have argued that every grammatical item displays a rich range of collocational patterns, from relatively variable collocational frameworks, to lexical phrases and fixed idiomatic expressions. These phrases in turn have patterns of phraseological use in the text which extend beyond the boundaries of the clause, an issue

which serves to enhance rather than distract from Halliday and Hasan's notion of textual cohesion.

It is worth admitting at this point that some features of phraseology which do not involve isolated grammatical items may have escaped our statistical trawling. It is fair to say that the reduced relative clauses mentioned in our sample Discussion section above would be missed by a preliminary analysis using *Wordlist*. Although reduced relatives involve a complex syntax and consistent morphology, this is one aspect of lexical collocation which is likely to be missed by our surface-based analysis. Generally speaking, there is no *a priori* reason why lexical collocations should not form part of the predominant phraseology of a textual genre. There is also no reason why morphological features of the text can not be taken into account. However, the fact remains that grammatical collocation is involved in an immense portion (if not a majority) of the typical kinds of expression to be found in a particular text.

These observations suggest that although collocational patterns must be an important first step in genre analysis, a closer reading of the text is also required. Typical grammatical phraseology clearly needs to be compared with other important lexical expressions. As we have seen in the sample text above, non-typical formulations are likely to have significant roles to play in the text. Another example from the corpus involves the unusual sentence adverb *'Forefront'* in the Introduction of Text *JNCI*: *Forefront in this role is tumor necrosis factor TNF...* Since the text is written by a native-speaker, it might be assumed that this is a rather marked expression, perhaps used to signal that this sentence, above all others, is worthy of notice (in popularised versions of this article *TNF* is hailed as a new discovery in our understanding of cancer, as we see below). Such interesting and significant features of the text should not be ignored, as they are also significant in terms of the text as a whole. But it is also clear that the idiosyncratic nature of individual texts can be only be demonstrated by establishing in the first instance those elements which are generic or salient in the broader corpus and ultimately in the general language as a whole.

Such exceptions to the rule also indicate that while the global analysis of collocation is essential in order to establish the major idiomatic characteristics of the corpus, statistical collocations can only be considered to be a limited area of style in which all the texts appear to overlap. Thus generic collocations are important in the sense that they lay bare those areas of the text which are truly individual or deviant. Such considerations have long been recognised in the statistical analysis of authorship (in science writing, Harris 1985), in forensic linguistics (Gibbons 1994) and studies on information retrieval (Sparck-Jones 1971, Choueka et al. 1985, Frohman

1990, Busch 1992). Once it is accepted that generic collocation is an important first step in describing the fundamental characteristics of a text, it appears increasingly unacceptable to adopt traditional approaches of literary analysis (and some discourse analysis), which stereotypically analyse the 'special' characteristics of a text without reference to a general phraseology of the genre, and ultimately of the language. In many ways this principle points out the insufficiency of my present study, and suggests that more related genres must be taken into account, such as a statistical comparison with a control corpus of general scientific texts and ultimately with a general corpus of English. This leads us naturally on to a discussion of future possibilities of research.

5. New Research Directions.

As I suggested in the previous section, the research set out in this book leaves a number of questions unanswered. It is not clear, for example, how phraseology in science is determined and propagated within the discourse community. There is no indication as yet whether the phraseological patterns we have seen in a very specific genre are replicated in disciplines other than cancer research. And there has been no space to discuss the historical dimension of phraseology. For example, a collocational account would certainly enhance the useful work carried out already by Biber and Finegan (1988) and Atkinson (1992) on the history of the research article genre. I have suggested above that the language of science can be defined in terms of mechanisms of reformulation and phraseology, in particular by the underlying tendency towards grammatical metaphor. But it must also be the case that the research article creates its own new phraseology, and that one aspect of successful research lies in the extent to which the new phraseology has been able to penetrate (or be accepted by) the existing discourse and be replicated as part of the established order. Studies such as Choueka et al. (1985) and Busch (1992) argue that slight variation in the use of common lexical collocations is an important indicator of novelty in technical writing. This suggests a future research programme which explores the possibility that language has a role to play in the natural selection of scientific ideas. I have previously proposed a phraseological view of logogenesis (the evolution of phrases within the text, Gledhill 1997), and would like to suggest that future work be applied to ontological development (the acquisition of phraseology in the individual) and phylogenic development (the evolution of phraseology over time).

Similarly, very little is known about the long-range cohesive functions of collocation. While rhetorical structure allows the reader to predict what is to be said on a broader scale, phraseological patterns may also be involved in what I term the indexical function of the scientific text. That is to say, the use of devices for browsing and skimming through a text. In their studies of signalling and use of rhetorical structure, Swales (1981), Nwogu (1989) and Sharp (1989) found that predictable elements of rhetorical structure and visual format help readers to identify which parts of the text to jump to, and to guess the content of conventional areas of the texts. But while such analysis helps to describe the linear reading of texts, it does not explain how scientists make a coherent account of a partially read text, or how parts of the text may be considered cohesive even at some distance apart, a notion that we have seen in the work of Hoey (1991). In the light of Dopkins and Morris's (1992) work on eye-fixation in reading, it may be possible to examine the extent to which collocations and other fixed expressions attract (or repulse) the reader's attention, thus having an important role in text processing. So in addition to key words, rhetorical structure and graphic format, it is worth considering whether grammatical parallelism, conventionalised phrases and cohesive networks might also be used as long range cohesive devices in the process of reading. Although work on the semiotics of non-verbal features of the scientific research article has recently been carried out by Tarasova (1993) and Lemke (1998), it may be worthwhile to examine the relationship between phraseology and the non-verbal features of scientific discourse.

Another fruitful area of research may lie in the phraseology of scientific popularisation. While there have been many studies of the popularisation of science (Nwogu and Bloor 1991, Myers 1991, Varttala 1999), few have concentrated on phraseology. Popularisation also constitutes a vast range of genres and text types, and extends beyond the stereotypical kind of text one normally associates with popular science (for example the scientific blockbuster, as explored by Fuller 1998). I have carried out a preliminary analysis of journalistic accounts of one of my expert informant's recent 'breakthroughs' (Gledhill forthcoming). As noted in section II.4, the Pharmaceutical Sciences department had a number of breakthroughs relating to the work of the microbiologist, *MT*. It turns out in fact that scientific breakthroughs are planned. The local and national press are informed at regular intervals of what to report and when. This degree of manipulation and interdependence between the press and the researchers changes our perspective on popularisation, and is interesting not in terms of the simplification of ideas, but in the way in which scientific discourse is used for rhetorical purposes.

It is possible to compare the phraseology of highly specialised texts such as *JNCI* with a corpus of articles such as the Daily Telegraph's *'Cancer discovery by farmer scientist*'. My initial findings suggest that popular accounts of scientific research are heavily influenced by the language of the scientists' reports. Interestingly, most reports devote only one or two lines to the actual 'science' of the story (the rest of the article concentrates on issues that are never dealt with in the research articles, such as the local angle and funding). When the press does explain the science, it appears that there is little effort to simplify the language involved. It is as though the journalist switches genres within the text. Here is the original formulation of the main scientific breakthrough from the *Biochemistry Journal:*

> The reason for depletion of host tissues is not known, but is thought to arise from differences in metabolism in the tumour-bearing state. (Biochemistry Journal)

From 12 newspaper clippings in the local and national press, the first sentence of the Independent suffices to show the processes of reformulation which may take place:

> A substance found in fish oil is to be used in the treatment of cancer, following new evidence that it can shrink solid tumours and may halt the dramatic weight loss associated with the disease. (The Independent)

The report displays several examples of phraseology which would not be out of place in the Pharmaceutical Sciences Corpus: nominal compaction (the use of *'of'* and reduced relative clauses) as well as hedging with *'may'*. In addition, there are a number of grammatical metaphors (underlined), expressing impersonal ideas (*treatment of..., new evidence that..., weight loss associated with...*).. There is therefore a striking similarity between this discourse and that of the original research articles. Since the journalists themselves use press releases produced by the cancer research charities, this is presumably reflected in the language of the popular report. Despite similar phraseological features, the press reports are never quite the same as each other, which leads to an interesting range of variable expressions. The consequences of this are not yet clear. But it would seem to suggest that stereotypical features of scientific writing such as nominalisation, passivisation and general complexity of grammatical metaphor are just as much a part of the popularised genre of science writing as the original technical text. Science writing becomes less bound to an original text or genre, and takes on a more abstract existence as a mode of meaning.

Beyond the corpus analysis carried out in this study, there is further work to be done in genre and discourse analysis in general. Despite the immense growth of specialised language corpora, there remains considerable scope for the analysis of collocation in both descriptive and applied linguistics. Very little work has been done for example on the comparative analysis of lexico-grammars in languages other than English. While much work in corpus linguistics has recently been devoted to language teaching (for example, Johns and King 1993, Van Halteren 1994), Barnbrook (1996) points out that corpora are long way from being properly exploited as reference tools in general linguistics. There is in contrast a strong tradition of corpus analysis in literary and authorship studies (more recently including Potter 1991 and Ide 1993) and there have been interesting developments in forensic linguistics and in the automatic detection of plagiarism (Coulthard 1994). But in each case there remains much to be said about the comparative analysis of collocation and phraseology. A large text corpus produced by second-language learners of English has been examined extensively by Granger (1996), and this research has shown that it is possible to examine collocational differences between apprentice writers and professionals in order to pin-point learners' difficulties and design teaching materials. A corpus of 'apprenticeship' texts may not only be a useful analytical tool in monitoring the linguistic progress of apprentice writers, but also in analysing how texts are edited and changed in their process of production, and how coherence develops chronologically throughout the text (such work has been taken on by Kouřilova, forthcoming). And in this respect, there are many dimensions of the Pharmaceutical Sciences Corpus which remain unexplored, for example the potential differences between single-author and team-authored texts, between native-speaker and non-native texts, or between papers on biology and those on structural chemistry. These fascinating possibilities belong, of course, to another book.

Language in Performance Series No. 22, Tübingen, Gunter Narr Verlag, 270pp.

VI. Appendix A: Frequency List.

The Most Frequent Words in the Pharmaceutical Sciences Corpus
(First 100 Items)[1].

1 THE	29122	(5.7%)	36 AFTER	1139	(0.2%)
2 OF	21309	(4.1%)	37 HAVE	1127	(0.2%)
3 AND	14610	(2.8%)	38 ML	1097	(0.2%)
4 IN	14349	(2.8%)	39 N (*nitrogen*)	1076	(0.2%)
5 TO	8631	(1.7%)	40 X (*algebraic*)	1045	(0.2%)
6 A (□)	8125	(1.6%)	41 IT	1006	(0.2%)
7 WAS	6146	(1.2%)	42 P (*pressure*)	992	(0.2%)
8 WITH	5543	(1.1%)	43 M (*mol./metre*)	973	(0.2%)
9 FOR	5224	(1.0%)	44 WE	972	(0.2%)
10 WERE	5162	(1.0%)	45 BEEN	966	(0.2%)
11 BY	4176	(0.8%)	46 TUMORS	903	(0.2%)
12 THAT	3352	(0.6%)	47 MICE	902	(0.2%)
13 AT	3287	(0.6%)	48 ALSO	884	(0.2%)
14 IS	3169	(0.6%)	49 ACTIVITY	880	(0.2%)
15 AS	3061	(0.6%)	50 G (*gramme*)	878	(0.2%)
16 CELLS	3016	(0.6%)	51 THAN	822	(0.1%)
17 FROM	2982	(0.6%)	52 D (*deuterium*)	821	(0.1%)
18 C (*celsius*)	2303	(0.4%)	53 USED	790	(0.1%)
19 OR	2290	(0.4%)	54 HUMAN	784	(0.1%)
20 ON	2182	(0.4%)	55 ALL	783	(0.1%)
21 I (*iodine*)	2029	(0.4%)	56 BETWEEN	780	(0.1%)
22 THIS	1197	(0.4%)	57 DNA	778	(0.1%)
23 ET	1987	(0.4%)	58 TABLE	774	(0.1%)
24 H (*hydrogen*)	1961	(0.4%)	59 FIG	757	(0.1%)
25 AL	1933	(0.3%)	60 RESULTS	755	(0.1%)
26 ARE	1920	(0.3%)	61 USING	752	(0.1%)
27 CELL	1905	(0.3%)	62 PROTEIN	751	(0.1%)
28 BE	1825	(0.3%)	63 HAS	741	(0.1%)
29 NOT	1798	(0.3%)	64 SHOWN	731	(0.1%)
30 AN	1438	(0.3%)	65 MIN	725	(0.1%)
31 WHICH	1422	(0.3%)	66 DATA	715	(0.1%)
32 THESE	1392	(0.3%)	67 BOTH	713	(0.1%)
33 L (*liquid*)	1299	(0.2%)	68 GROWTH	707	(0.1%)
34 TUMOR	1235	(0.2%)	69 OBSERVED	703	(0.1%)
35 S (*seconds*)	1203	(0.2%)	70 STUDY	701	(0.1%)

[1] Single letters (e.g. C, I, H) are left in the count as many of these represent chemical or mathematical symbols. There is some ambiguity over 'A' which may in some cases represent a determiner, the symbol 'α', or the symbol 'A' for relative atomic mass. 'I' always represents iodine, or 'electric current' or some mathematical variable in this corpus.

71 NO	694	(0.1%)	86 MORE	612	(0.1%)
72 B (□)	683	(0.1%)	87 ONLY	611	(0.1%)
73 ANALYSIS	682	(0.1%)	88 T (*time / temp*)	609	(0.1%)
74 TWO	682	(0.1%)	89 TREATMENT	606	(0.1%)
75 OTHER	673	(0.1%)	90 GROUP	599	(0.1%)
76 BUT	663	(0.1%)	91 EACH	595	(0.1%)
77 MAY	658	(0.1%)	92 PATIENTS	584	(0.1%)
78 FOUND	651	(0.1%)	93 DOSE	582	(0.1%)
79 FIGURE	650	(0.1%)	94 EXPRESSION	582	(0.1%)
80 EFFECT	649	(0.1%)	95 TIME	578	(0.1%)
81 OBTAINED	640	(0.1%)	96 LINES	573	(0.1%)
82 NORMAL	629	(0.1%)	97 HOWEVER	561	(0.1%)
83 E (*emf*)	623	(0.1%)	98 GENE	557	(0.1%)
84 ONE	619	(0.1%)	99 CONTROL	548	(0.1%)
85 MG	618	(0.1%)	100 MM	540	(0.1%)

VII. Appendix B: Texts Used in the PSC

The Pharmaceutical Sciences Corpus (PSC)
Reference Lists.

Journals are alphabetically listed according to the Science Citation Index mnemonic code (*CCP, CL* etc) and not according to title. The Journal's rank in the SCI (1988) impact factor table (compared with 1000 other journals) is listed as an approximate indicator of prestige. The relative size of the journal as a percentage of the corpus is also noted. A Unix-based word count has been used for this list, where the total corpus is of 150 papers, and 519 201 running words. For each paper one of several field classifications is noted (generally: cancer research / medicinal chemistry / pharmacology / structural chemistry). Only asterisked authors (usually the lead writer) are noted in the case of multiple author papers.

A.C. - Angewandte Chimie.
[SCI 1988 Rank=93 Corpus %=0.49]

AC: The Self-assembly of catenated cyclodextrins. [Supramolecular chemistry]
Author: DA, JS Source: author's ms, forthcoming

B.J. - Biochemistry Journal.
[SCI 1988 Rank=152 Corpus %=0.45]

BJ: Metabolic substrate utilization by tumour and host tissues in cancer cachexia. [Cancer Histopathology]
Author: MT. Source: Biochem J 277/371 1991

B.J.C. - British Journal of Cancer.
[SCI 1988 Rank=340 Corpus %=5.5]

BJC1:The influence of the schedule and the dose of gemcitabine on the anti- tumour efficacy in experimental human cancer [Cancer Chemotherapy] Author: TB. Source: Brit J. Can 68/1 1993

BJC2:Regulation of cytochrome P450 gene expression in human colon and breast tumour xenografts [Carcinogenesis] Author: MP, JR. Source: Brit J. Can 65/4 1992

BJC3: Allele loss from 5q21 (APCIMCC) and 18q21 (DCC) and DCC mRNA expression in breast cancer [Carcinogenesis] Author: GH Source: Brit J. Can 65/5 1992

BJC4:Comparative radioimmunotherapy using intact or F(ab')2 fragments of 13lI anti- CEA antibody in a colonic xenograft model [Cancer Radioimmunology] Author: FS. Source: Brit J. Can 65/6 1992

BJC5:Characterization of n-inedsine-resistant human sarcomas. [Cancer Chemotherapy]
Author: ML, OD,YD. Source: Brit J. Can 65/7 1992

BJC6:Strong HLA-DR expression in large bowel carcinomas is associated with good prognosis [Etiology/Histopathology] Author: CV, NB, OP. Source: Brit J. Can 65/8 1992

BJC7:Response to adjuvant chemotherapy in primary breast cancer: no correlation with expression of glutathione S-transferases [Cancer Chemotherapy] Author: AL. Source: Brit J. Can 68/3 1993

BJC8:pS2 is an independent factor of good prognosis in primary breast cancer [Etiology/Oncology]
Author: HT. Source: Brit J. Can 68/4 1993
BJC9:Serum pituitary and sex steroid hormone levels in the etiology of prostatic cancer - a population-based case-control study [Cancer Etiology/ Case study] Author: WP, IT, PL. Source: Brit J. Can 68/5 1993
BJC10:Expression of group-II phospholipase A2 in malignant and non-malignant human gastric mucosa [Cancer Immunohistochemistry] Author: WI. Source: Brit J. Can 68/7 1993
BJC11:Endogenous cortisol exerts antiemetic effect similar to that of exogenous corticosteroid [Chemotherapy] Author: CY. Source: Brit J. Can 68/9 1993

B.J.P- British Journal of Pharmacology.
[SCI 1988 Rank=84 Corpus %= 1.89]
BJP1:Antiarrhythmic drugs, clofilium and cibenzoline are potent inhibitors of glibenclamide-sensitive K+ currents in Xenopus oocytes [Pharmacology] Author: TH. Source: B.J. Phar 2/109/3 1991
BJP2: Attenuation of contractions to acetylcholine in canine bronchi by anendogenous nitric oxide-like substance [Pharmacology] Author: AG. Source: B.J. Phar 4/109/3 1991
BJP3: Enhancement by endothelin-1 of microvascular permeability via the activation of ETA receptors. [Pharmacology] Author: MT et al. . Source: B.J. Phar 5/109/3 1991

B.M.J. - British Medical Journal.
[SCI 1988 Rank=232 Corpus %=2.153]
BMJ1: The Bristol third stage trial: active versus physiological management of third stage of labour [Physiological management] Source:Astec corpus
BMJ2:Immunity to rubella in women of childbearing age in the United Kingdom [Etiology/Virology]
Source: Astec corpus
BMJ3:Adverse neurodevelopmental outcome of moderate neonatal hypoglycaemia [Physiological management] Source: Astec corpus
BMJ4:Seasonal distribution in conceptions achieved by artificial insemination by donor [Etiology/Gynacology] Source: Astec corpus
BMJ5: Aspirin and bleeding peptic ulcers in the elderly [Pharmacology] Source: Astec corpus

CAR - Carcinogenesis.
[SCI 1988 Rank=326 Corpus %=8.475]
CAR1:Sensitivity to tumor promotion of SENCAR and C57BL/6J mice correlates with oxidative events and DNA damage. [Tumour Promotor Carcinogenesis] Author: NH. Car. 4/5 1993
CAR2: Ras protooncogene activation of methylene chloride. [Carcinogenesis]
Author: CK. Car. 5/5 1993
CAR3:Characterization of p53 mutations in methylene chloride-induced lung tumors from B6C3F1 mice [Cancer Histology] Author: NE. Car. 1/6 1993
CAR4:Inhalation exposure to a hepatocarcinogenic concentration of methylene chloride does not induce sustained replicative DNA synthesis in hepatocytes of female B6C3F1 mice [Cancer Histopathology] Author: RS. Car. 2/6 1993

CAR5:Effect of varying exposure regimens on methylene chloride-induced lung and liver tumors in female B6C3F1 mice. [Chemical Carcinogenesis] Author: FP. Car. 3/6 1993

CAR6:Expression and stability of p53 protein in normal human mammary epithelial cells. [Tumour Supressor Gene Carcinogenesis] Author: GP. Car. 1/3 1992

CAR7: p53 Mutations in human immortalized epithelial cell lines [Carcinogenesis]
Author: YU. Car. 2/3 1992

CAR8:Protection against N-nitrosodiethylamine and benzo[a]pyrene-induced forestomach and lung tumorigenesis in A/J mice by green tea. [Cancer Immunohistochemistry] Author: LG. Car. 3/3 1992

CAR9 Inhibitory effects of curcumin on protein kinase C activity induced by 12-0-tetradecanoyl-phorbol-13-acetate in NIH 3T3 cells. [Cancer Immunohistochemistry] Author: MH. Car. 4/3 1992

CAR10 Characterization of highly polar bis- dihydrodiol epoxide-DNA adducts formed after metabolic activation of dibenz[a,h]anthracene [Carcinogenesis] Author: PR. Car. 5/3 1992

C.C. - Chemical Communications.
[SCI 1988 Rank=360 Corpus %=0.698]

CC: Bioreversible Protection for the Phospho Group: Chemical Stability and Bioactivation of Di(4-acetoxybenzyl) Methylphosphonate with Carboxyesterase [Structural chemistry] Author: SF, WJ, AM, DN, WT. J Chem Soc. 13/ 1991

C.C.P. - Cancer Chemotherapy and Pharmacology.
[SCI 1988 Rank=160 Corpus %=11.816]

CCP1:Quantification of the synergistic interaction of edatrexate and cisplatin in vitro. [Cancer Chemotherapy] Author: MP. 31/4 1993

CCP2 Pharmacokinetics of peptichemio in myeloma patients: release of m-L-sarcolysin in vivo and in vitro. [Cancer Chemotherapy] Author: CP. 31/5 1993

CCP3:Prolonged retention of high concentrations of 5-fluorouracil in human and murine tumors as compared with plasma. [Cancer Chemotherapy] Author: MP 31/6 1993

CCP4:Relationship between the melanin content of a human melanoma cell line and its radiosensitivity and uptake of pimonidazole. [Cancer Radioimmunology] Author:YW,PS 30/2 1992

CCP5:Phase I clinical and pharmacology study of 502U83 given as a 24-h continuous intravenous infusion. [Cancer Chemotherapy] Author: DD. 30/6 1992

CCP6:Correlation of the in vitro cytotoxicity of ethyldeshydroxysparsomycin and cisplatin with tne in vivo antitumour activity in murine L121~) leukaemia and two resistant L1210 subclones. [Cancer Chemotherapy]
Author: EL. 30/4 1992

CCP7:Doxorubicin and local hyperthermia in the microcirculation of skeletal muscle. [Cancer Chemotherapy] Author: AM. 30/3 1992

CCP8:Decreased resistance to N,N-dimethylated anthracyclines in multidrug-resistant Friend erythroleukemia cells. [Cancer Chemotherapy] Author: FJ. 30/1 1992

CCP9:Antitumor activity of the aromatase inhibitor FCE 24928 on DMBA-induced mammary tumors in ovariectomized rats treated with testosterone. [Cancer Chemotherapy] Author: IY. 29/6 1992

CCP10: Organ distribution and antitumor activity of free and liposomal doxorubicin injected into the hepatic artery [Cancer Chemotherapy] Author: DJ. 29/5 1992

CCP11: Effect of toremifene on antipyrine elimination in the isolated perfused rat liver.
Author: TD 29/4 1992
CCP12:A limited sampling method for estimation of the carboplatin area under the HNR curve. Cell-growth inhibition by and cytotoxicity of anthracyclines in doxorubicin-sensitive and -resistant F4-6 cells. [Cancer Chemotherapy] Author: PI. 29/3 1992
CCP13:Pharmacokinetics of 10-ethyl-10-deaza- aminopterin, edatrexate, given weekly for non- small-cell lung cancer [Cancer Chemotherapy] Author: KH. 29/2 1992
CCP14:Phase I clinical evaluation of [SP-4-3(R)]-[1,1-cyclobutanedicarboxylato(2-)] (2-methyl-1,4-butanediamine-N,Nl) platinum in patients with metastatic solid tumors [Cancer Chemotherapy] Author: VE. 29/1 1992
CCP15:Phase II study of high-dose ifosfamide in hepatocellular carcinoma [Cancer Chemotherapy]
Author: RW. 28/6 1992
CCP16: Ifosfamide in advanced epidermoid head and neck cancer [Cancer Chemotherapy]
Author: SI. 28/5 1992

C.L.- Cancer Letters.
[SCI 1988 Rank=251 Corps %=5.643]

CL1:Purification and analysis of a human sarcoma associated antigen [Cancer Chemotherapy]
Author: SG. 151/216 1 / 1993
CL2:Potentiation of butyrate-induced differentiation in human colon tumor cells by deoxycholate [Cancer Chemotherapy] Author: FT. 151/200 / 1993
CL3:Serum cross-reactive thymosin al levels in rats during induction of mammary carcinoma with 7,12-dimethylbenz[a]anthracene: short- and long-term effects. [Cancer Carcinogenesis] Author: KT. 151/218 / 1993
CL4:In vitro effects of natural plant polyphenols on the proliferation of normal and abnormal human lymphocytes and their secretions of interleukin-2 [Cancer Chemotherapy] Author: TU. 151/219 / 1993
CL5: Inhibition of melanoma cell growth by amino acid alcohols. [Cancer Chemotherapy] Author: RT 151/220 / 1993
CL6:p53 Mutations are common in pancreatic cancer and are absent in chronic pancreatitis [Carcinogenesis] Author: AS. 151/222/ 1993
CL7:Effect of exogenous heparin on anchorage-independent growth of fibroblasts induced by transforming cytokines [Cancer Immunohistochemistry] Author: HY. 151/203 / 1993
CL8:c-Ha-Ras mutants with point mutations in Gln-Val-Val region have reduced inhibitory activity toward cathepsin B [Cancer Immunohistochemistry] Author: HD. 151/204/ 1993
CL9:Inhibition of benzoyl peroxide-induced tumor promotion and progression by copper(II) (3,5-diisopropylsalicylate)2 [Cancer Carcinogenesis] Author: RS. 151/205 / 1993

C.R. - Cancer Research.
[SCI 1988 Rank=132 Corpus %=5.461]
CR1:Intracellular Localization of Human DNA Repair Enzyme Methylguanine-DNA Methyltransferase by Antibodies and its Importance. [Oncology] Author: IG Vol 53/21 1992

CR2: Monoclonal Antibodies to the Myogenic Regulatory Protein MyoD1: Epitope Mapping and Diagnostic Utility. [Cancer Immunohistochemistry] Author: TW Vol 53/23 1992

CR3:Therapy with Unlabeled and 13lI-labeled Pan-B-Cell Monoclonal Antibodies in Nude Mice Bearing Raji Burkitt's Lymphoma Xenografts [Cancer Immunohistochemistry] Author: ET Vol 53/24 1992

CR4: Inhibition of Cellular Proliferation by Peptide Analogues of Insulin-like Growth Factor [Cancer Chemotherapy] Author: LK Vol 53/25 1992

CR5:Expression of the Endogenous 06-Methylguanine-DNA-methyltransferase Protects Chinese Hamster Ovary Cells from Spontaneous G:C to A:T Transitions1 [Cancer Carcinogenesis] Author: PS Vol 54/26 1993

CR6:Tumor-associated Mr 34,000 and Mr 32,000 Membrane Glycoproteins That Are Serine-Phosphorylated Specifically in Bovine Leukemia Virus-induced Lymphosarcoma Cells' [Cancer Carcinogenesis] Author:PR Vol 54/27 1993

CR7:Antitumor Effect of Interferon plus Cyclosporine A following Chemotherapy for Disseminated Melanomal [Cancer Immunology] Author: SH Vol 54/28 1993

CR8: Tumorigenic Suppression of a Human Cutaneous Squamous Cell Carcinoma Cell Line in the Nude Mouse Skin Graft Assay. [Cancer chemotherapy] Author: GU Vol 54/29 1993

CR9:A Retrovirus in Chinook Salmon (OncoYhynchus tshawytscha) with Plasmacytoid Leukemia and Evidence for the Etiology of the Disease. [Carcinogenesis] Author: AL Vol 52/17 1991

CR10: Expression and CpG Methylation of the Insulin-like Growth Factor II Gene in Human Smooth Muscle Tumors [Carcinogenesis] Author: HT Vol 52/18 1991

CR11:Loss of Heterozygosity Involves Multiple Tumor Suppressor Genes in Human Esophageal Cancers [Carcinogenesis] Author: YF Vol 54/19 1991

CR12:Induction of c-fos Gene Expression by Exposure to a Static Magnetic Field in HeLaS3 Cells1 [Carcinogenesis] Author: KH Vol 54/20 1991

F.A.T. - Fundamental and Applied Toxicology.
[SCI 1988 Rank= 289 Corpus %=7.3]

FAT1:2,4,5-Trichlorophenoxyacetic Acid Influence on 2,6-Dinitrotoluene induced Urine Genotoxicity in Fischer 344 Rats: Effect on Gastrointestinal Microflora and Enzyme Activity [Toxicology] Author BN. Source F. App. Tox. 18/2 1992

FAT2:Three-Month Effects of MDL 19,660 on the Canine Platelet and Erythrocyte [Toxicology] Author IY. Source F. App. Tox. 18/3 1992

FAT3:Evaluation of the Potential for Developmental Toxicity in Rats and Mice following Inhalation Exposure to Tetrahydrofuran [Toxicology] Author GH. Source F. App. Tox. 18/3 1992

FAT4:Topical Anesthetic-lnduced Methemoglobinemia in Sheep: A Comparison of Benzocaine and Lidocaine1. [Toxicology] Author PK. Source F. App. Tox. 18/4 1992

FAT5: Time Course of Permeability Changes and PMN Flux in Rat Trachea following 03 Exposure [Toxicology] Author JG. Source F. App. Tox. 19/1 1993

FAT6:Control of the Nephrotoxicity of Cisplatin by Clinically Used Sulfur-Containing Compounds [Toxicology] Author LW. Source F. App. Tox. 19/2 1993

FAT7: Developmental Toxicity of Boric Acid in Mice and Rats. [Toxicology] Author FG. Source F. App. Tox. 19/3 1993

FAT8:Acrylamide: Dermal Exposure Produces Genetic Damage in Male Mouse Germ Cells. [Toxicology] Author GN. Source F. App. Tox. 19/4 1993

FAT9: Effects of Diet Type on Incidence of Spontaneous and 2-Acetylaminofluorene-lnduced Liver and Bladder Tumors in BALB/c Mice Fed AIN-76A Diet versus NIH-07 Diet [Toxicology] Author PO. Source F. App. Tox. 17/ 1 1991

FAT10: Risk Assesment in Immunotoxicity. Sensitivity and Predictability of Immune Tests. [Toxicology] Author SA. Source F. App. Tox. 17/3 1991

IJ.C.- International Journal of Cancer.
[SCI 1988 Rank= 226 Corpus %= 17.556]

IJC1:Down-regulation of ri(x) subunit of camp-dependent protein kinase induces growth inhibition of human mammary epithelial cells transformed by c-ha-ras and c-erbb-2 proto-oncogenes [Cancer Cytogenetics] Author: TM. Source: Int J. Cancer 53/14 1992

IJC2:Phenotypic and molecular analysis of ph-chromosome-positive acute lymphoblastic leukemia cells. [Cancer Cytogenetics] Author: . Source: Int J. Cancer 53/72 1993

IJC3: Loss of heterozygosity at the short arm of chromosome 3 in renal-cell cancer correlates with the cytological tumour type [Cancer Cytogenetics] Author: AH et al.. Source: Int J. Cancer 53/61 1992

IJC4:Over-expression of p53 nuclear oncoprotein in transitional-cell bladder cancer and its prognostic value [Cancer Cytogenetics]. Author: PL. Source: Int J. Cancer 53/62 1992

IJC5:International variations in the incidence of childhood bone tumours [Cancer Epidemiology]
Author: DP, CS, JN. Source: Int J. Cancer 53/63 1992

IJC6:Molecular and serological studies of human papillomavirus among patients with anal epidermoid carcinoma [Cancer Epidemiology] Author: PH, SG, UL, JD. Source: Int J. Cancer 53/64 1992

IJC7:Concordant p53 and dcc alterations and allelic losses on chromosomes 13q and 14q associated with liver metastases of colorectal carcinoma [Cytogenetics] Author: KO et al. Source: Int J. Cancer 53/66 1992

IJC8: Isolation and characterization of an oestrogen- responsive breast-cancer cell line, eff-3 [Cancer Cytogenetics] Author: RH et al. Source: Int J. Cancer 53/671992

IJC9:Differential regulation of gelatinase b and tissue-type plasminogen activator expression in human Bowes melanoma cells [Cancer Histopathology] Author: HB, RZ. Source: Int J. Cancer 53/68 1992

IJC10:Antibody-induced growth inhibition is mediated through immunochemically and functionally distinct epitopes on the extracellular domain of the c-erbb-2 (her-2/neu) gene product pl85 [Cancer Immunohistochemistry] Author: FX et al. Source: Int J. Cancer 53/69 1992

IJC11: Structure-activity relationships of four anti-cancer alkylphosphocholine derivatives in vitro and in vivo [Cancer Chemotherapy]. Author: SS et al. . Source: Int J. Cancer 53/70 1992

IJC12:Analysis of the relationship between stage of differentiation and NK/LAK susceptibility of colon carcinoma cells. [Cancer Histopathology] Author: HB, RZ. Source: Int J. Cancer 53/72 1993

IJC13:Combination effect of vaccination with il2 and il4 cdna transfected cells on the induction of a therapeutic immune response against lewis lung carcinoma cells [Cancer Cytogenetics] Author: YO, EP,KO. Source: Int J. Cancer 53/74 1993

IJC14: Comparative cytogenetic and dna flow cytometric analysis of 150 bone and soft-tissue tumors [Cytogenetics] Author: NM, BB etc.. Source: Int J. Cancer 53/84 1993

IJC15: The role of the urokinase receptor in extracellular matrix degradation by ht29 human colon carcinoma cells [Cancer Histopathology] Author: LR, EK. Source: Int J. Cancer 53/85 1993

IJC16: Immortalization of normal human fibroblasts by treatment with 4-nitroquinoline l-oxide. [Cancer Cytogenetics] Author: LB, YK, MN. Source: Int J. Cancer 53/86 1993

IJC17:Expression and distribution of peripherin protein in human neuroblastoma cell lines. [Cancer Histopathology] Author: HB, RZ. Source: Int J. Cancer 53/87 1993

IJC18:Anti-metastatic vaccination of tumor-bearing mice with il-2-gene-inserted tumor cells. [Cancer Immunohistochemistry] Author: AP, BG,RB. Source: Int J. Cancer 53/88 1993

IJC19: Distinct p-glycoprotein expression in two subclones simultaneously selected from a human colon carcinoma cell line by cis-diamminedichloroplatinum (ii) [Cancer Chemotherapy] Author: LY, JT. Source: Int J. Cancer 53/89 1993

IJC20:Cellular and in vivo characterization of the mcr rat mammary tumor model [Cancer Immunohistochemistry] Author: AG, UR. Source: Int J. Cancer 53/90 1993

IJC21:Co-amplification of c-myc/pvt-l in immortalized mouse b-lymphocytic cell lines results in a novel pvt-l/aj-l transcript. [Cytogenetics] Author: KH, DS. Source: Int J. Cancer 53/91 1993

IJC22:Persistence of plasmin-mediated pro-urokinase activation on the surface of human monocytoid leukemia cells in vitro. [Cancer Histopathology] Author: HT. Source: Int J. Cancer 53/92 1993

IJC23:Cytokeratins expressed in experimental rat bronchial carcinomas [Cancer Histopathology]

Author: HK, AHB etc.. Source: Int J. Cancer 53/93 1993

IJC24:Activators of coagulation in cultured human lung-tumor cells [Cancer Histopathology]

Author: RS, HH. Source: Int J. Cancer 53/94 1993

IJC25:Action of a cd24-specific deglycosylated ricin-a-chain immunotoxin in conventional and novel models of small-cell-lung-cancer xenograft. [Cancer Immunohistochemistry] Author: UP, HPL. Source: Int J. Cancer 53/95 1993

J.C.P.T. - Journal of Chemistry: Perkin Transactions.

[SCI 1988 Rank= 290 Corpus %= 6.626]

JCPT1: Synthesis of (+)- and (-)-Methyl Shikimate from Benzene [Structural Chemistry]

Author CJ Vol 1 1993

JCPT2: A Reinvestigation of the Intramolecular Buchner Reaction of 1- Diazo-4-phenylbutan-2-ones Leading to 2-Tetralones [Structural Chemistry] Author AC Vol 2 1993

JCPT3:Synthesis of ‘5N-Labelled Chiral Boc-Amino Acids from Triflates of Leucine and Phenylalanine. [Structural Chemistry] Author FD Vol 3 1993

JCPT4:Studies on Pyrazines. Part 25. Lewis Acid-promoted Deoxidative Thiation of Pyrazine N-Oxides: New Protocol for the Synthesis of 3-Substituted Pyrazinethiols. [Structural Chemistry] Author NS Vol 4 1993

JCPT5: Use of the 1-(2-Fluorophenyl)-4-methoxypiperidin-4-yl (Fpmp) Protecting Group in the Solid-Phase Synthesis of Oligo- and Poly-ribonucleotides. [Structural Chemistry] Author VR Vol 4 1992

JCPT6:Reinvestigation of the Pummerer Arylation of to 2,2’,5’-Trihydroxybiaryls. Quinones: A Selective Approach. [Structural Chemistry] Author GS Vol 2 1992

JCPT7:Synthesis and Hydrolysis Studies of Phosphonopyruvate. [Structural Chemistry] Author: SF Vol. 2 1991

JCPT8:Structural Studies on Bio-active Molecules. Part 17. Crystal Structure of 9-(2'-Phosphonylmethoxyethyl)adenine (PMEA). [Structural Chemistry]. Authors: WT, SF. Source: author ms

JCPT9:Bioreversible Protection for the Phospho Group: Bioactivation of the Di(4-acyloxybenzyl) and Mono(4-acyloxybenzyl) Phosphoesters of Methylphosphonate and Phosphonoacetate1. [Structural Chemistry] Author: AM, WT, DN, WI, SF.Vol 1 1992

JCPT10:Latent Inhibitors. Part 7. Inhibition of Dihydro-orotate Dehydrogenase by Spirocyclopropanobarbiturates. [Structural Chemistry].. Author: WF, CS, HW 1 1990

J.G.M. - Journal of General Microbiology.
[SCI 1988 Rank= 389 Corpus %= 7.971]

JGM1: Isolation and characterization of urease from AspeYgillus niger. [Enzymology]
Author RD. JGM Vol 193/5 1992

JGM2:Functional and physiological characterization of the Tn21 cassette for resistance genes in Tn2426 [Enzymology] Author JG. JGM Vol 193/8 1992

JGM3:Resistance to spiramycin in Streptomyces ambofaciens, the producer organism involves at least two different mechanisms. [Enzymology] Author SJ. JGM Vol 189/1 1989

JGM4:The induction of oxidative enzymes in Streptomyces coelicolor upon hydrogen peroxide treatment. [Enzymology] Author PF. JGM Vol 189/2 1989

JGM5:Bacterial metabolism of 5-aminosalicylic acid: enzymic conversion to L-malate, pyruvate and ammonia. [Enzymology] Author SK. JGM Vol 189/3 1989

JGM6:Regulation of methylthioribose kinase by methionine in Klebsiella pneumoniae. [Enzymology].
Author ME. JGM Vol 189/4 1989

JGM7:Ionophoric action of trans-isohumulone on Lactobacillus brevis. [Immunobacteriology]
Author BU. JGM Vol 190/2 1990

JGM8:Archetal halophins (halobacteria) from 2 salt enzymes in klebsiella pneumoniae. [Enzymology]
Author BI. JGM Vol 190/3 1990

JGM9: Characterization of the trypsin-like enzymes of Polyphyomonas gingivalis W83 using a radiolabelled active-site-directed inhibitor. [Enzymology] Author LD. JGM Vol 188/1 1988

J.M.C. - Journal of Medicincal Chemistry.
[SCI 1988 Rank= 384 Corpus %= 0.86]

JMC: Structural Studies on Tazobactam. [Structural Chemistry]
Author PL. J MedChem 34 / 1991

J.N.C.I. - Journal of the National Cancer Institute.
[SCI 1988 Rank= Not ranked. Corpus %= 0.39]

JNCI: Lipolytic Factors Associated With Murine and Human Cancer Cachexia [Cancer Histopathology]
Author HD, MT. JNat Can Inst 82/24 1990

J.O.A.C.S. - Journal of the American Chemical Society.
[SCI 1988 Rank= 312. Corpus %= 6.179]

Language in Performance Series No. 22, Tübingen, Gunter Narr Verlag, 270pp.

JOACS1:Time Evolution of the Intermediates Formed in the Reaction of Oxygen with Mixed-Valence Cytochrome c Oxidase. [Sructural Chemistry] Author: WH JOrgS. Vol. 112/26 1991

JOACS2:Dynamic Properties and Electrostatic Potential Surface of Neutral DNA Heteropolymers. [Organic Chemistry] Author: SN JOrgS. Vol. 112/25 1991

JOACS3:Bonding between C2 and N2: A Localization- Induced (a) Bond. [Organic Chemistry]

Author: KL JOrgS. Vol. 112/27 1991

JOACS4:Normal-Mode Characteristics of Chlorophyll Models. Vibrational Analysis of Metallooctaethylchlorins and Their Selectively Deuterated Analogues. [Organic Chemistry]

Author: AD JOrgS. Vol. 112/16 1991

JOACS5:The Effect of §-Fluorine Substituents on the Rate and Equilibrium Constants for the Reactions of ~-Substituted 4-Methoxybenzyl Carbocations and on the Reactivity of a Simple Quinone Methide. [Organic Chemistry] Author: MK JOrgS. Vol. 113/9 1992

JOACS6:Concurrent Stepwise and Concerted Substitution Reactions of 4-Methoxybenzyl Derivatives and the Lifetime of the 4-Methoxybenzyl Carbocation. [Structural Chemistry] Author: NE JOrgS. Vol. 113/6 1992

JOACS7: Enzyme and mediated enantiface differentiation.[Organic Chemistry] Author: SC JOrgS. Vol. 113/7 1992

JOACS8:Photochemical Ligand Loss as a Basis for Imaging and Microstructure Formation in a Thin Polymeric Film. [Structural Chemistry] Author: VN JOrgS. Vol. 113/8 1992

JOACS9:IHNMR Resonance Assignment of the Active Site Residues of Paramagnetic Proteins by 2D Bond Correlation Spectroscopy: Metcyanomyoglobin. [Organic Chemistry] Author: BN JOrgS. Vol. 113/10 1992

JOACS10 How Far Can a Carbanion Delocalize? 13C NMR Studies on Soliton Model Compounds. [Organic Chemistry] Author: WA JOrgS. Vol. 113/11 1992

JOACS11:Calculation of Structures and Bond Dissociation Energies of Radical Cations: The Importance of Through-Bond Delocalization in Bibenzylic Systems.[Organic Chemistry] Author: SG JOrgS. Vol. 114/1 1993

J.O.C. - Journal of Organic Chemistry.
[SCI 1988 Rank= 382 Corpus %= 5.940]

JOC1:Oxidation of Natural Targets by Dioxiranes. 2.1 Direct Hydroxylation at the Side- Chain C-25 of Cholestane Derivative and of Vitamin D3 Windaus-Grundmann Ketone. [Organic Chemistry] Author LE: JOC 57/6 1992
JOC2:Synthesis of 3-Arylpyrroles and 3-Pyrrolylacetylenes by Palladium-Catalyzed Coupling Reactions [Organic Chemistry] Author JH: JOC 57/5 1992
JOC3:A Simple Asymmetric Synthesis of 2-Substituted Pyrrolidines and 5-Substituted Pyrrolidinones [Organic Chemistry] Author MR: JOC 57/4 1992
JOC4:Stereo-and Regioselective Synthesis Of Chiral Diamines and Triamine from Pseudoephedrine and Ephedrine [Organic Chemistry] Author PD: JOC 57/1 1992
JOC5: New Electron Acceptors: Synthesis, Electrochemistry, and Radical Anions of N,7,7-Tricyanoquinomethanimines and X-ray Crystal Structures of the Trimethyl and Tetramethyl Derivatives [Organic Chemistry] Author IS: JOC 57/2 1992
JOC6:Stereocontrolled Syntheses of Substituted Unsaturated Lactam from 3-Alkenamide [Organic Chemistry] Author ST: JOC 57/3 1992
JOC7: Importance of the Folded Orientation of Two Enoate Moietiey [Organic Chemistry]
Author: FN JOC 58/1 1993

J.P.P.- Journal of Pharmacy and Pharmacology.
[SCI 1988 Rank= 465 Corpus %= 3.195]

JPP1:Hydrolysis of Partially Saturated Egg Phosphatidylcholine in Aqueous Liposome Dispersions and the Effect of Cholesterol Incorporation on Hydrolysis Kinetics [Pharmacology] Author RY, SJ, HS: JPP 46/6 1990
JPP2:Hydrolysis and Stability of Acetylsalicylic Acid in Stearylamine-containing Liposomes [Pharmacology] Author: DI, SA, IS JPP 46/5 1990
JPP3: In-vitro Bioadhesion of a Buccal, Miconazole Slow-release Tablet [Pharmacology]
Author RT, SG: JPP 46/4 1990

P.A.H. - Pharmaceutica Acta Helvetica.
[SCI 1988 Rank= 516. Corpus %= 0.726]

PAH1:Thin Layer Chromatography in Pharmaceutical Quality Control. Assay of Inosiplex in different pharmaceutical forms. [Pharmacology] Author ED: Pharm A Helv 67/342-373
PAH2:The Stability of Famotidine Hydrochloride Solutions at Different pH Values. [Pharmacology]
Author LK: Pharm A Helv 67/321-352

Language in Performance Series No. 22, Tübingen, Gunter Narr Verlag, 270pp.

T.L. - Tetrahedron Letters.
[SCI 1988 Rank= 476. Corpus %=0.446]

TL: Synthesis of Antiviral Nucleosides from Crotonaldehyde. Part 3.1,2 Total Synthesis of Didehydrodideoxythymidine (d4T) [Organic Chemistry] Author: JE, JG. Tetr Let Vol. 33/27 1992

T.P.S. - Trends in Pharmaceutical Sciences.
[SCI 1988 Rank= 94. Corpus %=0.231]

TPS: Newly identified factors that alter host metabolism in cancer cachexia [Cancer Histopathology]
Author: MT. Source: JNCI Vol. 82/ 24

VIII. Appendix C: Salient Word Lists

1. Salient Words in Titles[2]

RANK	*WORD*	*Titles Freq.*	*%*	*PSC Freq.*	*%*	*Chi²*	*Probability*
1	CHARACTERIZATI	8	(0.4%)	44		236.0	
2	HUMAN	25	(1.2%)	784	(0.2%)	126.6	
3	SYNTHESIS	12	(0.6%)	204		119.9	
4	LNDUCED [sic]	2		3		101.4	
5	KLEBSIELLA	2		4		84.0	
6	REINVESTIGATIO	2		4		84.0	
7	METHOXYBENZYL	3	(0.1%)	14		80.3	
8	CANCER	16	(0.7%)	522	(0.1%)	74.8	
9	METHYLTRANSFER	2		5		71.6	
10	EDATREXATE	2		5		71.6	
11	CARCINOMA	9	(0.4%)	205		62.2	
12	OF	166	(7.6%)	21309	(4.3%)	59.3	0.000
13	BIOREVERSIBLE	2		7		55.0	
14	13LI	2		8		49.2	
15	B6C3F1	3	(0.1%)	24		48.8	
16	SUBSTITUTES	5	(0.2%)	77		48.6	
17	METHYLGUANINE	2		10		40.5	
18	EXPRESSION	13	(0.6%)	582	(0.1%)	38.4	
19	EPIDERMOID	2		12		34.3	
20	PNEUMONIAE	2		13		31.8	
21	REGULATION	4	(0.2%)	72		30.7	
22	N	17	(0.8%)	1076	(0.2%)	29.4	
23	LEUKEMIA	4	(0.2%)	75		29.3	
24	FLUX	1		1		28.0	
25	L121	1		1		28.0	
26	VLVO [sic]	1		1		28.0	
27	POLYPHYOMONAS	1		1		28.0	
28	E1	1		1		28.0	
29	AMINOSALICYLIC	1		1		28.0	
30	SERINEPHOSPHOR	1		1		28.0	
31	LIDOCAINE1	1		1		28.0	
32	ONCOYHYNCHUS	1		1		28.0	

[2] Some items were mis-scanned in the original corpus. I have marked them *sic*

33	INEDSINE	1	1	28.0
34	MELANOMAL	1	1	28.0
35	MOIETIEY	1	1	28.0
36	SUBLCONES [sic]	1	1	28.0
37	ASSAY1	1	1	28.0
38	LYMPHOBLASTIC	1	1	28.0
39	AANALYSIS [sic]	1	1	28.0
40	PYRENEINDUCED	1	1	28.0
41	ARCHETAL	1	1	28.0
42	IMPORTANCEL	1	1	28.0
43	ANTLTUMOUR [sic]	1	1	28.0
44	ASPEYGILLUS	1	1	28.0
45	DISEASE1	1	1	28.0
46	DELOCALIZE	1	1	28.0
47	PREDICTABILITY	1	1	28.0
48	TRIAMINE	1	1	28.0
49	PREDICTABILITY	1	1	28.0
50	TRIAMINE	1	1	28.0

Salient Grammatical Words in Titles

		Titles		*PSC*			
RANK	***WORD***	***Freq.***	***%***	***Freq.***	***%***	***Chi²***	***Probability***
12	OF	166	(7.6%)	21309	(4.3%)	59.3	0.000
60	FOR	110	(5.0%)	5224	(1.0%)	26.6	0.000
67	ON	24	(1.1%)	2182	(0.4%)	20.5	0.000
70	AND	99	(4.6%)	14610	(2.9%)	19.7	0.000
134	IN	91	(4.2%)	14349	(2.9%)	12.9	0.000

2. Salient Words in Abstracts

		Abstracts		*PSC*			
RANK	***WORD***	***Freq.***	***%***	***Freq.***	***%***	***Chi²***	***Probability***
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 [sic]	4		4		22.4	
18	SPIRAMYEIN	4		4		22.4	
19	TUMOR	114	(0.4%)	1235	(0.2%)	21.8	0.000
20	INHIBITED	21		121		21.7	0.000
21	IOA	6		12		21.7	
22	EXPRESSION	63	(0.2%)	582	(0.1%)	21.6	0.000
23	PATIENTS	63	(0.2%)	584	(0.1%)	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%)	521	(0.1%)	20.1	0.000
29	OA	4		5		19.2	
30	NDENT	5		9		19.0	
31	BUT	67	(0.2%)	663	(0.1%)	18.1	0.000
32	IMMORTALIZED	13		62		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	(0.1%)	16.4	0.000
38	DECREASED	23		161		15.9	0.000
39	CANCER	54	(0.2%)	522	(0.1%)	15.7	0.000
40	CONTRACTIONS	5		11		15.7	

41	AZIDE	10		43		15.7	
42	HAEMORRHAGE	8		29		15.5	
43	THESE	119	(0.4%)	1399	(0.3%)	15.3	0.000
44	MANAGEMENT	17		104		15.3	0.000
45	ETHOXY	3		3		15.0	
46	PROFICIENT	3		3		15.0	
47	NONNAL	3		3		15.0	
48	BENZOCAINE	12		61		14.7	
49	PAA	4		7		14.6	
50	TUMORS	82	(0.3%)	903	(0.2%)	14.4	0.000

Salient Grammatical Words in Abstracts

		Abstracts		*PSC*			
RANK	***WORD***	***Freq.***	***%***	***Freq.***	***%***	***Chi²***	***Probability***
31	BUT	67	(0.2%)	663	(0.1%)	18.1	0.000
43	THESE	119	(0.4%)	1399	(0.3%)	15.3	0.000
79	OF	1367	(4.7%)	21309	(4.3%)	11.8	0.001
198	THERE	40	(0.1%)	444		6.5	0.011
203	IN	912	(3.1%)	14349	(2.9%)	6.3	0.012
267	WAS	365	(1.3%)	6271	(1.2%)	5.0	0.020
299	THAT	227	(0.8%)	3357	(0.7%)	4.5	0.034
329	DID	34	(0.1%)	395		4.3	0.037
334	WHO	14		129		4.2	0.040
378	BOTH	55	(0.2%)	713	(0.1%)	3.7	0.055

3. Salient Words in Introduction Sections

		Introductions		*PSC*			
RANK	***WORD***	***Freq.***	***%***	***Freq.***	***%***	***Chi²***	***Probability***
1	ET	692	(1.2%)	1987	(0.4%)	652.5	0.000
2	AL	670	(1.1%)	1933	(0.4%)	626.3	0.000
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
6	INTRODUCTION	83	(0.1%)	97		234.8	0.000
7	IS	643	(1.1%)	3169	(0.6%)	156.3	0.000
8	RECENTLY	52		102		84.3	0.000
9	STUDIES	135	(0.2%)	494		76.6	0.000
10	CANCER	140	(0.2%)	522	(0.1%)	76.0	0.000
11	SUCH	113	(0.2%)	388		73.7	0.000
12	GENES	82	(0.1%)	242		71.9	0.000
13	EFFECTS	112	(0.2%)	414		61.8	0.000
14	VARIETY	37		72		59.9	0.000
15	CAN	120	(0.2%)	468		58.1	0.000
16	ROLE	56		152		56.4	0.000
17	REPORT	37		79		53.0	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
20	SUPPRESSOR	39		92		48.5	0.000
21	HUMAN	167	(0.3%)	784	(0.2%)	47.4	0.000
22	IMPORTANT	55		170		43.7	0.000
23	MANY	50		150		41.9	0.000
24	SYNTHESIS	61	(0.1%)	204		41.5	0.000
25	OF	2874	(4.8%)	21309	(4.3%)	41.4	0.000
26	CHIRAL	26		51		41.0	0.000
27	ARE	332	(0.6%)	1920	(0.4%)	39.7	0.000
28	BE	317	(0.5%)	1825	(0.4%)	38.8	0.000
29	SEVERAL	75	(0.1%)	284		38.7	0.000
30	REPORTED	95	(0.2%)	395		38.6	0.000
31	CLINICAL	48		151		36.7	0.000
32	TO	1233	(2.1%)	8631	(1.7%)	36.6	0.000
33	COMPOUNDS	76	(0.1%)	296		36.6	0.000
34	MECHANISMS	45		138		36.1	0.000
35	ITS	88	(0.1%)	365		36.0	0.000
36	OFTEN	29		68		35.9	0.000
37	SYSTEMS	37		104		34.5	0.000
38	CANCERS	36		100		34.3	0.000
39	SOME	77	(0.1%)	310		34.0	0.000

40	AGENTS	45		145		32.7	0.000
41	ACYLOXYMETHYL	1		11		31.9	
42	DEMONSTRATED	48		162		31.8	0.000
43	THIS	330	(0.6%)	1997	(0.4%)	30.6	0.000
44	USEFUL	26		63		30.4	0.000
45	PROPERTIES	28		73		29.3	0.000
46	GENE	115	(0.2%)	557	(0.1%)	29.0	0.000
47	ATTENTION	14		21		28.7	
48	VIVO	48		171		28.2	0.000
49	MAY	130	(0.2%)	658	(0.1%)	27.9	0.000
50	INCLUDE	21		47		27.2	0.000

Salient Grammatical Words in Introduction Sections.

		Introductions		***PSC***			
RANK	***WORD***	***Freq.***	***%***	***Freq.***	***%***	***Chi²***	***Probability***
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		73.7	0.000
15	CAN	120	(0.2%)	468		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%)	21309	(4.3%)	41.4	0.000
32	TO	1233	(2.1%)	8631	(1.7%)	36.6	0.000

4. Salient Words in Methods Sections.

		Methods		*PSC*			
RANK	***WORD***	***Freq.***	***%***	***Freq.***	***%***	***Chi²***	***Probability***
1	WERE	2795	(2.0%)	5162	(1.0%)	876.5	0.000
2	H	1281	(0.9%)	1961	(0.4%)	620.2	0.000
3	WAS	2877	(2.1%)	6146	(1.2%)	576.7	0.000
4	ML	850	(0.6%)	1097	(0.2%)	562.8	0.000
5	C	1303	(0.9%)	2303	(0.5%)	454.8	0.000
6	MIN	506	(0.4%)	725	(0.1%)	277.5	0.000
7	MM	401	(0.3%)	540	(0.1%)	245.9	0.000
8	MMOL	282	(0.2%)	302		245.4	0.000
9	ADDED	295	(0.2%)	340		231.6	0.000
10	M	582	(0.4%)	973	(0.2%)	231.2	0.000
11	X	597	(0.4%)	1045	(0.2%)	212.4	0.000
12	G	520	(0.4%)	878	(0.2%)	201.7	0.000
13	D	487	(0.4%)	821	(0.2%)	189.5	0.000
14	SOLUTION	304	(0.2%)	428		171.7	0.000
15	HZ	240	(0.2%)	294		171.5	0.000
16	S	620	(0.5%)	1203	(0.2%)	166.9	0.000
17	WASHED	179	(0.1%)	190		157.0	0.000
18	THEN	282	(0.2%)	420		142.9	0.000
19	BUFFER	232	(0.2%)	313		141.2	0.000
20	AT	1324	(1.0%)	3287	(0.7%)	140.3	0.000
21	PH	304	(0.2%)	483		134.8	0.000
22	USING	412	(0.3%)	752	(0.2%)	131.2	0.000
23	PBS	143	(0.1%)	153		123.8	0.000
24	INCUBATED	184	(0.1%)	237		120.9	0.000
25	FOR	1919	(1.4%)	5224	(1.0%)	120.1	0.000
26	DESCRIBED	269	(0.2%)	436		114.0	0.000
27	WATER	209	(0.2%)	305		109.9	0.000
28	PERFORMED	181	(0.1%)	250		105.3	0.000
29	SODIUM	142	(0.1%)	173		101.7	0.000
30	EACH	323	(0.2%)	595	(0.1%)	100.2	0.000
31	CONTAINING	229	(0.2%)	370		97.6	0.000
32	V	288	(0.2%)	515	(0.1%)	96.5	0.000
33	I	828	(0.6%)	2029	(0.4%)	93.1	0.000
34	USED	391	(0.3%)	790	(0.2%)	92.7	0.000
35	SIGMA	100		102		91.7	0.000
36	CH	100		106		87.2	0.000
37	COLUMN	152	(0.1%)	212		86.7	0.000
38	DRIED	102		113		83.7	0.000
39	MEDIUM	221	(0.2%)	376		83.6	0.000
40	DISSOLVED	90		92		82.1	0.000
41	TEMPERATURE	145	(0.1%)	204		81.3	0.000
42	MIXTURE	137		188		80.4	0.000
43	MHZ	92		101		76.3	0.000

44	AND	4633	(3.4%)	14610	(2.9%)	74.3	0.000
45	METHODS	162	(0.1%)	253		74.0	0.000
46	ROOM	99		117		73.9	0.000
47	CM3	81		84		72.4	0.000
48	DILUTED	79		82		70.5	0.000
49	COLLECTED	102		128		69.3	0.000
50	REMOVED	102		132		65.9	0.000

Salient Grammatical Words in Methods Sections.

		Methods		***PSC***			
RANK	***WORD***	***Freq.***	***%***	***Freq.***	***%***	***Chi²***	***Probability***
1	WERE	2795	(2.0%)	5162	(1.0%)	876.5	0.000
3	WAS	2877	(2.1%)	6146	(1.2%)	576.7	0.000
18	THEN	282	(0.2%)	420		142.9	0.000
20	AT	1324	(1.0%)	3287	(0.7%)	140.3	0.000
25	FOR	1919	(1.4%)	5224	(1.0%)	120.1	0.000
30	EACH	323	(0.2%)	595	(0.1%)	100.2	0.000
44	AND	4633	(3.4%)	14610	(2.9%)	74.3	0.000
82	FROM	1048	(0.8%)	2982	(0.6%)	47.2	0.000
139	AFTER	431	(0.3%)	1139	(0.2%)	32.0	0.000
260	WITH	1711	(1.2%)	5543	(1.1%)	17.8	0.000

5. Salient Words in Results Sections

RANK	*WORD*	*Results Freq.*	*%*	*PSC Freq.*	*%*	*Chi²*	*Probability*
1	FIGURE	470	(0.4%)	650	(0.1%)	366.3	0.000
2	FIG	496	(0.4%)	757	(0.2%)	328.1	0.000
3	TABLE	475	(0.4%)	774	(0.2%)	278.7	0.000
4	SHOWN	372	(0.3%)	731	(0.1%)	145.4	0.000
5	P	451	(0.4%)	992	(0.2%)	130.6	0.000
6	H69	126	(0.1%)	163		107.4	0.000
7	MEAN	207	(0.2%)	364		103.5	0.000
8	CELLS	1028	(0.9%)	3016	(0.6%)	95.7	0.000
9	VALUES	231	(0.2%)	453		90.3	0.000
10	TREATED	225	(0.2%)	449		84.2	0.000
11	LANE	142	(0.1%)	230		83.3	0.000
12	CONTROL	257	(0.2%)	548	(0.1%)	80.9	0.000
13	SPIRAMYCIN	98		136		74.7	0.000
14	LLC	118		184		74.1	0.000
15	SHOWS	121	(0.1%)	197		70.1	0.000
16	NO	296	(0.2%)	694	(0.1%)	70.0	0.000
17	OBSERVED	298	(0.2%)	703	(0.1%)	69.1	0.000
18	LANES	83		113		65.0	0.000
19	SIGNIFICANTLY	150	(0.1%)	291		59.9	0.000
20	KG	154	(0.1%)	303		59.4	0.000
21	D122	85		126		57.9	0.000
22	VDS	70		92		57.6	0.000
23	SIGNIFICANT	181	(0.2%)	386		56.7	0.000
24	ANIMALS	227	(0.2%)	524	(0.1%)	56.3	0.000
25	B	275	(0.2%)	683	(0.1%)	53.2	0.000
26	MYCELIUM	56		67		52.4	0.000
27	SHOWED	172	(0.1%)	375		50.5	0.000
28	IN	3906	(3.3%)	14349	(2.9%)	50.4	0.000
29	DID	176	(0.1%)	395		47.5	0.000
30	NOT	595	(0.5%)	1798	(0.4%)	46.5	0.000
31	NUB	52		65		45.6	0.000
32	DAYS	191	(0.2%)	446		45.5	0.000
33	LIVER	201	(0.2%)	479		44.8	0.000
34	VERAPAMIL	62		89		44.2	0.000
35	WEEKS	142	(0.1%)	304		43.8	0.000
36	COMPARED	162	(0.1%)	364		43.5	0.000
37	HAD	206	(0.2%)	517	(0.1%)	38.2	0.000
38	LINES	221	(0.2%)	573	(0.1%)	36.1	0.000
39	RESULTS	275	(0.2%)	755	(0.2%)	35.2	0.000
40	AJ	43		57		34.3	0.000
41	AFTER	385	(0.3%)	1139	(0.2%)	33.8	0.000
42	MRNA	103		215		33.8	0.000

43	LOH	104		218		33.8	0.000
44	MR	57		91		33.6	0.000
45	GROUPS	163	(0.1%)	397		33.6	0.000
46	TIME	219	(0.2%)	578	(0.1%)	33.3	0.000
47	LEVELS	192	(0.2%)	491		33.1	0.000
48	CODON	55		87		33.0	0.000
49	INCIDENCE	96		197		32.9	0.000
50	POSITIVE	124	(0.1%)	282		31.9	0.000

Salient Grammatical Words in Results Sections.

		Results		***PSC***			
RANK	***WORD***	***Freq.***	***%***	***Freq.***	***%***	***Chi²***	***Probability***
16	NO	296	(0.2%)	694	(0.1%)	70.0	0.000
28	IN	3906	(3.3%)	14349	(2.9%)	50.4	0.000
29	DID	176	(0.1%)	395		47.5	0.000
30	NOT	595	(0.5%)	1798	(0.4%)	46.5	0.000
37	HAD	206	(0.2%)	517	(0.1%)	38.2	0.000
41	AFTER	385	(0.3%)	1139	(0.2%)	33.8	0.000
72	THERE	168	(0.1%)	444		25.2	0.000
80	THE	7427	(6.2%)	29122	(5.8%)	23.4	0.000
92	WHEN	184	(0.2%)	518	(0.1%)	20.8	0.000
125	ALL	252	(0.2%)	783	(0.2%)	16.3	0.000

6. Salient Words in Discussion Sections.

		Discussions		*PSC*			
RANK	***WORD***	***Freq.***	***%***	***Freq.***	***%***	***Chi²***	***Probability***
1	THAT	1381	(1.2%)	3357	(0.7%)	341.8	0.000
2	BE	788	(0.7%)	1825	(0.4%)	225.6	0.000
3	MAY	383	(0.3%)	658	(0.1%)	223.2	0.000
4	IS	1167	(1.0%)	3169	(0.6%)	193.1	0.000
5	ET	789	(0.7%)	1987	(0.4%)	172.6	0.000
6	AL	762	(0.7%)	1933	(0.4%)	162.4	0.000
7	OUR	222	(0.2%)	381		129.0	0.000
8	DISCUSSION	119	(0.1%)	145		119.1	0.000
9	IN	3991	(3.5%)	14349	(2.9%)	116.0	0.000
10	MODES	131	(0.1%)	179		111.6	0.000
11	NOT	662	(0.6%)	1798	(0.4%)	108.9	0.000
12	THIS	704	(0.6%)	1997	(0.4%)	96.2	0.000
13	WE	395	(0.3%)	972	(0.2%)	92.9	0.000
14	HAVE	442	(0.4%)	1127	(0.2%)	92.1	0.000
15	STUDY	306	(0.3%)	701	(0.1%)	89.8	0.000
16	ENDOTHELIN	162	(0.1%)	303		78.6	0.000
17	IT	390	(0.3%)	1006	(0.2%)	77.8	0.000
18	MODE	91		136		66.9	0.000
19	P53	175	(0.2%)	376		61.0	0.000
20	PRESENT	189	(0.2%)	419		60.5	0.000
21	CAN	205	(0.2%)	468		60.5	0.000
22	MIGHT	110		196		58.7	0.000
23	SUGGEST	102		177		57.4	0.000
24	HOWEVER	231	(0.2%)	561	(0.1%)	56.4	0.000
25	HAS	285	(0.2%)	741	(0.1%)	55.1	0.000
26	REPORTED	176	(0.2%)	395		54.4	0.000
27	THESE	475	(0.4%)	1399	(0.3%)	54.1	0.000
28	COULD	176	(0.2%)	398		53.2	0.000
29	STRETCHING	59		78		51.9	0.000
30	FINDINGS	71		108		50.4	0.000
31	SUCH	166	(0.1%)	388		45.5	0.000
32	WHICH	468	(0.4%)	1422	(0.3%)	45.4	0.000
33	BEEN	339	(0.3%)	966	(0.2%)	45.0	0.000
34	THE	7292	(6.4%)	29122	(5.8%)	44.4	0.000
35	MORE	232	(0.2%)	612	(0.1%)	42.3	0.000
36	GENE	212	(0.2%)	557	(0.1%)	39.2	0.000
37	EXPRESSION	219	(0.2%)	582	(0.1%)	38.8	0.000
38	SUGGESTS	68		117		38.5	0.000
39	CUOEC	64		107		38.2	0.000
40	WOULD	108		232		37.3	0.000
41	DOES	67		117		36.8	0.000
42	INCREASE	144	(0.1%)	352		34.1	0.000

43	PROBABLY	58		101		31.9	0.000
44	SUGGESTED	59		104		31.7	0.000
45	PERMEABILITY	55		94		31.3	0.000
46	ARE	576	(0.5%)	1920	(0.4%)	31.2	0.000
47	INDICATE	77		155		31.1	0.000
48	MECHANISMS	71		138		31.1	0.000
49	TO	2261	(2.0%)	8631	(1.7%)	30.6	0.000
50	DUE	108		252		29.5	0.000

Salient Grammatical Words in Discussion Sections

		Discussions		***PSC***			
RANK	***WORD***	***Freq.***	***%***	***Freq.***	***%***	***Chi²***	***Probability***
1	THAT	1381	(1.2%)	3357	(0.7%)	341.8	0.000
2	BE	788	(0.7%)	1825	(0.4%)	225.6	0.000
3	MAY	383	(0.3%)	658	(0.1%)	223.2	0.000
4	IS	1167	(1.0%)	3169	(0.6%)	193.1	0.000
7	OUR	222	(0.2%)	381		129.0	0.000
9	IN	3991	(3.5%)	14349	(2.9%)	116.0	0.000
11	NOT	662	(0.6%)	1798	(0.4%)	108.9	0.000
12	THIS	704	(0.6%)	1997	(0.4%)	96.2	0.000
13	WE	395	(0.3%)	972	(0.2%)	92.9	0.000
14	HAVE	442	(0.4%)	1127	(0.2%)	92.1	0.000

IX. References

Aarts Jan. 1992. 'Comments' in J. Svartvik 1992a: 180-183

Aarts J. and Meijs W. (eds.) 1984 *Corpus Linguistics. Recent Developments in the Use of Corpora in English Language Research* Amsterdam: Rodopi

Aarts J. and Meijs W. (eds.) 1986 *Corpus Linguistics II* Amsterdam: Rodopi

Aarts J. and Meijs W. (eds.) 1990 *Theory and Practice in Corpus Linguistics* Amsterdam: Rodopi

Abeillé A. 1995. 'The Flexibility of French Idioms: A Representation with Lexicalized Tree Adjoining Grammar.' in M. Everaert et al. (eds.): 15-42

Abraham E. 1991. 'Why 'Because'? the Management of Given / New Information as a Constraint on the Selection of Causal Alternatives.' in *Text* Vol.11/3: 323-339

Adams-Smith D.E. 1984. 'Medical Discourse: Aspects of Authors' Comments.' in *English for Specific Purposes Journal* Vol.3/1: 25-36

Adams-Smith D.E. 1987. 'Variation in Field-Related Genres.' in *English Language Research Journal* Vol.1: 10-32

Ager D.E. 1976. 'The Importance of the Word in the Analysis of Register.' in A. Jones and R.F. Churchhouse (eds.) *The Computer in Linguistic and Literary Studies*, University of Wales Press: 55-68

Ager D.E., Knowles F.E. and J. Smith 1979 (eds.). *Advances in Computer-Aided Literary and Linguistic Research* Birmingham: Aston University

Ahmad K., Fulford H., Griffin S. and Holmes-Higgins 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.

Alexander R. J. 1978. 'Fixed Expressions in English: Reference Books and the Teacher' in *English Language Teaching Journal*. 38/2: 127-134.

Alexander R. J. 1989. 'Fixed Expressions, Idioms and Collocations Revisited.' in P. Meara (ed.) *Beyond Words. British Studies in Applied Linguistics 4.* Proceedings of B.A.A.L'98, Exeter, September 1988. Pp15-25.

Alexander R. J. 1991. 'Hopes and Fears of a Corpus Linguist or, the Sad but Edifying Tale of A Corpus Search for Fixed Expressions.' in *Corpora des Englischen in Forschung, Lehre und Anwendungen* (CCE Newsletter) Vol. 5 (1/2): 1- 12

Altenberg B. 1991. 'Amplifier Collocations in Spoken English.' in S. Johansson and A.B. Stenström (eds.) 1991: 127-147

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. 1990. 'Register: A Review of Empirical Research.' in D. Biber and E. Finegan (eds.) 1991b: 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

Auger C.P. 1989. *Information Sources in Grey Literature* London: Bowker-Saur

Austin J.L. 1962 * 1975 (eds. Urmson J.O. and Sbisà M). *How to Do Things with Words* London: Oxford University Press

Baker D.B., Horiszy J.W. and Metanomski W.V. 1980. 'History of Abstracting at Chemical Abstracts Service.' in *Journal of Chemical Information and Computer Science* Vol. 20: 193-201

Baker M., Francis G. and Tognini-Bonelli E. (eds.) 1993. *Text and Technology* Amsterdam: John Benjamins

Barber C.L. 1962. 'Some Measurable Characteristics of Modern Scientific Prose.' in Almquist and Wikwell (eds.) *Contributions to English Syntax and Philology*: 21-43

Barnbrook G. 1996. *Language and Computers* Edinburgh University Press: Edinburgh

Barthes R. 1966. *Mythologies*. Paris: Seuil

Basili R., Pazienza M.T. and Velardi P. 1992. 'A Shallow Syntactic Analyser to Extract Word Associations from Corpora.' in *Literary and Linguistic Computing* Vol.7/2: 113-123

Banks D. 1994a 'Clause Organization in the Scientific Journal Article'. *Alsed-Lsp Newsletter* Vol. 17/2: 4-16.

Banks D 1994b. *Writ in Water: Aspects of the Scientific Journal Article*. E.R.L.A.: Université De Bretagne.

Banks D. 1997. 'The Things We Make'. In *Language Sciences*. 19/4: 303-308.

Banks D. 1998. 'Vague Quantification in the Scientific Journal Article.' in *Anglais de Spécialité.* GERAS: Presses de l'Université Victor-Segalen, Bordeaux No. 19/22: 17-27.

Bauer L. 1979. 'On the Need for Pragmatics in the Study of Nominal Compounding.' in *Journal of Pragmatics*. 3/1: 45-50.

Béjoint H. 1988. 'Scientific and Technical Words in General Dictionaries.' in *International Journal of Lexicography* Vol. 1/4: 354-368

Benson M. 1989. 'The Collocational Dictionary and the Advanced Learner.' in M.L. Tickoo (ed.) *Learner's Dictionaries: State of the Art* Singapore: SEAMO Regional Language Centre: 84-93

Benson. M., Benson., E. and Ilson R. 1986 *The Lexicographic Description of English* London: John Benjamins

Bernier C.L. 1972. 'Terse Literatures 1: Terse Conclusions.' in *Journal of the American Society for Information Science* Vol. 21: 316-319

Bernier C.L. 1985. 'Abstracts and Abstracting.' in *DYM*: 423-444

Berry-Rogghe G. 1970. 'Collocations: Their Computation and Semantic Significance.' Unpublished Ph.D Thesis, UMIST, Manchester

Biber D.1986. *Variation across Speech and Writing* Cambridge: Cambridge University Press

Biber D. 1989. 'A Typology of English Texts.' in *Linguistics* 27: 3-43

Biber D. 1992a. 'On the Complexity of Discourse Complexity: A Multidimensional Analysis.' in *Discourse Processes* Vol. 15 133-163

Biber D. 1992b. 'Using Computer-Based Text Corpora to Analyze the Referential Strategies of Spoken and Written Texts.' in J. Svartvik (ed.) 1992: 215-252

Biber D. 1993. 'The Multidimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings.' in *Computers and the Humanities* Vol. 26: 331-345.

Biber D. Conrad S. and Reppen R. 1994. 'Corpus-Based Approaches to Issues in Applied Linguistics.' in *Applied Linguistics* Vol. 15/2: 169-189

Biber D., Conrad S., and Reppen R. 1996. 'Corpus-Based Investigations of Language Use'. In *Annual Review of Applied Linguistics*. 16: 115-136.

Biber D., Conrad S., Reppen R. 1998. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Biber D. and Finegan E. 1988. 'Drift in Three English Genres from the 18th to the 20th Centuries: A Metadiscoursal Approach.' in M.Kytö et al. (eds.): 83-99

Biber D. and Finegan E. (eds.) 1994. *Sociolinguistic Perspectives on Register* Oxford: Oxford University Press

Blackwell S. 1987. 'Problems in the Automatic Parsing of Idioms.' in R. Garside et al. (eds) *Syntax Versus Orthography*: 110-119

Bloor T. and Bloor M. 1985. 'Language for Specific Purposes: Practice and Theory'. CLCS Occasional Papers: Trinity College, Dublin.

Borko H. and Chatman S. 1963. 'Criteria for Acceptable Abstracts: A Survey of Abstractors' Instructions.' in *American Documentation* Vol. 14: 175-184

Boyer E. 1994. *the Academic Profession: An International Perspective*. California: Princeton Press

Brekke M. 1991. 'Automatic Parsing Meets the Wall.' in S. Johansson and A.B. Strenström (eds.): 83-103

Brett P. 1994. 'A Genre Analysis of the Results Sections of Sociology Articles.' in *English for Specific Purposes Journal* Vol.13/1: 47-59

Briscoe T. 1990 'English Noun-Phrases Are Regular: A Reply to Professor Sampson.' in J. Aarts and W. Meijs 1990: 45-60

Britt M.A. Perfetti C.A. and Garrod S. 1992. 'Parsing in Discourse: Context Effects and Their Limits.' in *Journal of Memory and Language* Vol.31: 293-314

Burnard L. 1992. 'Tools and Techniques for Computer-Aided Text Processing.' in C. Butler (ed.): 1-28

Busch G. 1992. 'Search and Retrieval.' in *BYTE*, June: 274-282. New Yorx: Bix Publishers

Butler C. 1985a. *Computers in Linguistics* Oxford: Basil Blackwell

Butler C. 1985b. *Statistics in Linguistics* Oxford: Basil Blackwell

Butler C. (ed.) 1992. *Computers and Written Texts* Oxford: Basil Blackwell

Butler C. 1993. 'Between Grammar and Lexis: Collocational Frameworks in Spanish' Unpublished Paper Presented at the 5th International Systemic Workshop on Corpus-Based Studies, Universidad Complutense De Madrid, 26-29 July 1993

Buxton A.B. and Meadows A.J. 1978. 'Categorisation of Information in Experimental Papers and Their Author Abstracts.' in *Journal of Research Communication Studies* Vol. 1: 161-182

Cahn, R. S. 1979. *Introduction to Chemical Nomenclature*. New York Press.

Carter R. 1998. *Vocabulary. Applied Linguistic Perspectives.* (2nd Edition). London: Routledge.

Cavalli-Sforza L. and Felman M. 1989. *Cultural Transmission and Evolution* Princeton New Jersey: Princeton University Press

Chafe W. 1992. 'The Importance of Corpus Linguistics to Understanding the Nature of Language.' in Svartvik 1992a: 79-97

Chesterman A. 1997. *Memes of Translation. the Spread of Ideas in Translation Theory.* Amsterdam: John Benjamins.

Choueka Y., Klein T. and Neuwitch E. 1983. "'Automatic Retrieval of Idiomatic and Collocational Expressions in A Large Corpus.' in *Journal for Literary and Linguistic Computing* Vol. 4: 34-38

Church K. W. and Hanks . P 1989. 'Word Association Norms, Mutual Information and Lexicography.' in *Computational Linguistics* 16/1: 22-29

Church K. W. and Mercer R.L. 1993. 'Introduction to the Special Issue on Computational Linguistics Using Large Corpora.' in *Computational Linguistics* Vol. 19/1: 1-24

Clarke D. F. and Nation I. S. P. 1980. 'Guessing the Meanings of Words from Context: Strategy and Techniques'. In *System* 8/3: 211-220.

Clear J. 1987. 'Overview of the Role of Computing in Cobuild.' in J.McH. Sinclair (ed.) 1987: 41-61
Clear J. 1993. 'from Firth Principles. Collocational Tools for the Study of Collocation.' in M. Baker et al. (eds.) 1993: 271-292
Cleveland D.B. and Cleveland A.D. 1983. *Introduction to Indexing and Abstracting* Princeton Colorado Libraries Unlimited
Collins P. and Peters P. 1988. 'The Australian Corpus Project.' in M. Kytö et al. (eds.): 103-120
Collot M. 1991. 'Electronic Language. A Pilot Study of A New Variety of English. *Computer Corpora des Englischen in Forschung, Lehre und Anwendungen* (CCE Newsletter, Berlin) Vol. 5 (1/2): 13-31
Coulmas F. 1979. 'On the Sociolinguistic Relevance of Routine Formulae'. In *Journal of Pragmatics*. 2/3: 223-235.
Coulthard M. (ed.) 1994. *Advances in Written Text Analysis* London: Routledge.
Cowie A.P. (ed). 1998 *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press.
Cremmins E.T. 1982. T*he Art of Abstracting* Philadelphia ISI Press
Cruse D.A. 1986. *Lexical Semantics* Cambridge University Press
De Beaugrande R. 1991. *Linguistic Theory*. *the Discourse of Fundamental Works*. Longman: London.
De Beaugrande R. and Dressler W. 1981. *Introduction to Text Linguistics* London: Longman
DeCarrico, J. and Nattinger, J. 1988. `Lexical Phrases for the Comprehension of Academic Lectures', ESP Journal, 7/2, 91-101
Derewianka B. 1994. 'Grammatical Metaphor and Fuzzy Boundaries'. Unpublished MS, Presented at the 21st International Systemic Functional Comgress, 1-5 August 1994.
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
Dobrovol'skij, D. 1992. 'Phraseological Universals: Theoretical and Applied Aspects'. In M. Kefer (ed.) *Meaning and Grammar: Cross-Linguistic Perspectives*. Berlin.
Dopkins S. and Morris R.K. 1992. 'Lexical Ambiguity and Eye Fixation in Reading: A Test of Competing Models of Lexical Autonomy Resolution.' in *Journal of Memory and Language* Vol.31: 461-476
Dronberger G.B. and Kronitz G.T. 1975 'Abstract Readability As A Factor in Information Systems.' in *Journal of the American Society for Information Science* Vol. 26: 108-111
Drury H. 1991. 'The Use of Systemic Linguistics to Describe Student Summaries at University Level.' in E. Ventola (ed.) 1991: 431-456
Dubois B. L. 1981. 'The Construction of Noun Phrases in Biomedical Journal Articles.' in J. Hoedt et al. (eds) *Pragmatics and LSP* Copenhagen: : 49-67
Dubois B. L. 1997. *The Biomedical Discussion Section in Context*. London: Ablex Publishing Corporation.
Endres-Niggemeyer B. 1985. 'Referierregeln Und Referate- Abstracting Als Regelgesteuerter Textverarbeitungsprozeß.' in *Nachtrichten Für Dokumentaristen* Vol. 36/1: 38-50
Enkvist N. 1964. 'On Defining Style: An Essay in Applied Linguistics.' in J. Spencer (ed.) *Linguistics and Style* London: Oxford University Press.
Enkvist N. 1989. 'From Text to Interpretability: A Contribution to the Discussion of Basic Terms in Text Linguistics.' in W. Hyedrich et al. (eds.) 1989: 369-382
Escarpit R. 1976. *Théorie Générale de l'Information et de la Communication* Paris Hachette

Everaert M., Van Der Linden E., Schenk A., and Schreuder R. (eds.) 1995. *Idioms: Structural and Psychological Perspectives*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Fernando C. 1996. *Idioms and Idiomaticity*. Oxford: Oxford University Press.
Fidel R. 1986. 'Writing Abstracts for Free-Text Searching.' in *Journal of Documentation* Vol. 42/1: 11-21
Fillmore C.J. 1992. 'Corpus Linguistics, or Computer-Aided Armchair Linguistics.' in Svartvik (ed) 1992a: 35-60
Fillmore C.J. and Atkins S. 1994. 'Starting Where the Dictionaries Stop: the Challenge of Corpus Lexicography.' in S. Atkins and Zampolli (eds.) *Computational Approaches to the Lexicon* Oxford: Oxford University Press
Fillmore C.J., Kay P. and O'Connor M.C. 1988. 'Regularity and Idiomacy in Grammatical Constructions.' in *Language* Vol. 64: 501-538
Firth J.R. 1935. 'The Techniques of Semantics.' in *Transactions of the Philological Society*. 36-72.
Firth J.R. 1957. *Papers in Linguistics 1934-1951*. Oxford: Oxford University Press.
Fischer R. 1998. *Lexical Change in Present-Day English*. Tübingen: Gunter Narr Verlag.
Fløttum K. 1985. 'Methodological Problems in the Analysis of Student Summaries.' in *Text* Vol. 5/4: 291-308
Fontenelle T. 1994. 'What on Earth are Collocations?'. In *English Today* No. 40 Vol. 10/4: 42-48.
Fox G. 1993. 'A Comparison of 'Policespeak' and 'Normalspeak': A Preliminary Study.' in J. McH. Sinclair et al. (eds.) 1993: 184-195
Foucauld M. 1972. *the Archaeology of Knowledge* London: Tavistock.
Francis G. 1985. 'Anaphoric Nouns.' Discourse Analysis Monograph No. 11: Birmingham: Birmingham University English Language Research
Francis G. 1993. 'A Corpus-Driven Approach to Grammar.' in Baker et al. (eds.) 1993: 137-156
Francis G. and Kramer-Dahl A. 1991. 'From Clinical Report to Clinical Story: Two Ways of Writing About A Medical Case.' in E. Ventola (ed.) 1991: 339-368
Francis G. and SINCLAIR J. 1994. 'I Bet He Drinks Carling Black Label. A Riposte to Owen on Corpus Grammar.' in *Applied Linguistics* Vol.15/2: 188-200
Fuller G. 'Cultivating Science: Negotiating Discourse in the Popular Texts of Stephen Jay Gould'. In J. R. Martin , R. Veel (eds). 1998. *Reading Science: Critical and Functional Perspectives on Discourses of Science*. London: Routledge. 35-62.
Gadamer H.G. 1976. 'On the Scope and Function of Hermeneutical Reflection.' in D.E. Linge (ed. and Trans.) *Philosophical Hermeneutics* University of California Press.
Gerbert M. 1970. *Besonderheiten der Syntax in der Technischen Fachsprache des Englischen* Berlin: Halle.
Gerson S. 1989. '*From ...to* as an Intensifying Collocation.' in *English Studies* Vol. 70: 360-371
Gibson T.R. 1992. 'Towards a Discourse Theory of Abstracts and Abstracting.' Unpublished Ph.D. Thesis, English Language Department: Nottingham
Gibbons J. 1994. *Language and the Law.* London: Addison Wesley.
Gläser, R. 1989. 'Gibt Es Eine Fachsprachenphraseologie?', in Fachsprache - Fremdsprache - Muttersprache, VIIth International Conference 'Angewandte Sprachwissenschaft Und Fachsprachliche Ausbildung': Technische Universität Dresden
Gläser R. 1991. 'The LSP Genre Abstract - Revisited.' in *ALSED - Newsletter* Vol. 13/4: 3-11
Gläser R. 1992. 'A Multi-Level Model for a Typology of LSP Genres.' in *Fachsprache* Vol. 15/1-2: 18-26

Gläser R. 1998. 'The Stylistic Potential of Phraseological Units in the Light of Genre Analysis'. In A. P. Cowie (ed.): 125-143.

Gledhill C. 1995a. 'Collocation and Genre Analysis. ' In *Zeitschrift für Anglistik und Amerikanistik* Vol. 1:11-36

Gledhill C. 1995b. 'Scientific Innovation and the Phraseology of Rhetoric. Posture, Reformulation and Collocation in Cancer Research Articles.' PhD thesis, University of Aston.

Gledhill C. 1996. 'Science as a Collocation. Phraseology in Cancer Research Articles'. in Botley S., Glass J, McEnery T and Wilson A (eds.) *Proceedings of Teaching and Language Corpora 1996*. Lancaster. UCREL Technical Papers Volume 9: 108-126.

Gledhill C. 1997. 'Les collocations et la construction du savoir scientifique.' in Martin J. *Anglais de Spécialité (ASp).* No. 15-18 :85-104.

Gledhill C. 1999. 'Towards a phraseology of English and French'. In C. Beedham (ed.) *Language and Parole in Synchronic and Diachronic Perspective*. Proceedings of Societas Linguistica Europaea XXXI. Oxford: Pergamon: 221-37.

Gledhill C. (forthcoming) 'The phraseology of rhetoric, collocations and discourse in cancer research abstracts' in C. Barron and N. Bruce (eds.) . *'Knowledge and Discourse' Proceedings of the International Multidisciplinary Conference*. Hong Kong: 18-21 June 1996. University of Hong Kong, Hong Kong. April 1999.

Gnutzmann L. and Oldenburg H. 1992. 'Contrastive Text Linguistics in LSP Research: Theoretical Considerations and Some Preliminary Findings.' in Schneider (ed.): 103-136

Godley T. 1993. 'Terminological Principles and Methods in the Subject Field of Chemistry' in B. Sonneveld and Loening (eds.): 141-163

Godman A. and Payne E.M.F. 1981 'A Taxonomic Approach to the Lexis of Science.' in Selinker et al. (eds.) 23-39

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

Grätz N 1985. 'Teaching EFL Students to Extract Structural Information from Abstracts.' in J.M. Kline and A.K. Pugh (eds.) *Reading for Professional Purposes: Methods and Materials in Teaching Languages*: 225-335

Granger S. 1998. 'Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae'. In Cowie A. (ed) 1998: 1-21.

Grice H.P. 1975. 'Logic and Conversation' in P. Cole and J.Morgan (eds.) *Syntax and Semantics III* New York: Academic Press

Guba E.G. and Lincoln Y.S. 1982. 'Epistemological and Methodological 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* Vol. 8/3: 265-273.

Halliday MA.K. 1961. *Categories of the Theory of Grammar*. Department of English Language and General Linguistics Monographs. (Pp241-292). Edinburgh: Edinburgh University Press.

Halliday M.A.K. 1966. 'Lexis As A Linguistic Level' in Bazell et al. (eds.) 1966 *in Memory of J.R.Firth* London: Longman

Halliday M.A.K. 1976. 'Functions and Universals of Language.' in G. Kress (ed.) 1976 *Halliday: System and Function in Language* London: Oxford University Press

Halliday M.A.K. 1977. 'Language Structure and Language Function.' in J.Lyons (ed.) 1977 *New Horizons in Linguistics* Harmonsworth: Penguin Books

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

Halliday M.A.K. 1988. 'On the Language of Physical Science.. In M.Ghadessy 1988: 162-177

Halliday M.A.K. 1991a. 'Corpus Studies in Probabilistic Grammar.' in K. Aijmer and B. Altenberg (eds) 1991: 30-43
Halliday M.A.K. 1991b. 'Towards Probabilistic Interpretations.' in E. Ventola (ed.) 1991: 39-61
Halliday M.A.K. 1992. 'Language as System and Language As Instance: the Corpus As A Theoretical Construct.' in J. Svartvik (ed.) 1992a: 61-77
Halliday M.A.K. 1994. 'The Construction of Knowledge and Value in the Grammar of Scientific Discourse, with Reference to Charles Darwin's 'The Origin of Species'.' in M. Coulthard (ed.): 136-156.
Halliday M. A K. 1998. 'Things and Relations. Regrammaticising Experience as Technical Knowledge.' in J. R. Martin , R. Veel (eds) 1998 *Reading Science: Critical and Functional Perspectives on Discourses of Science*. London: Routledge. 185-235.
Halliday M.A.K. and James Z.L. 1993. 'A Quantitative Study of Polarity and Primary Tense in the English Finite Clause.' in J. McH. Sinclair (et al.) 1993: 32-66
Halliday M.A.K. and Hasan R. 1976. *Cohesion in English* London: Longman
Halliday M.A.K. and Hasan R. 1989. (2nd Edition) *Language, Context and Text: Aspects of Language in a Social-Semiotic Perspective* Oxford: Oxford University Press
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 M.Sc. Theses in Biology, Chemistry, and Physics.' in *English for Specific Purposes* Vol. 4: 49-58
Harley B. 1996. *Lexical Issues in Language Learning*. London: John Benjamins.
Harris J. E. 1985. 'Aspects of Authorship in the Scientific Abstract.' Unpublished MSc. Dissertation, Language Studies Unit: Aston University
Hartley J. 1994. 'Three Ways to Improve the Clarity of Journal Abstracts' in *British Journal of Educational Psychology* Vol. 64/2: 331-343
Heidegger M. 1966. *Discourse on Thinking* London: Torch: Harper and Row
Hoey M. 1983. *On the Surface of Discourse* London: Allen and Unwin
Hoey M. 1991. *Patterns of Lexis in Text* Oxford: Oxford University Press
Hopkins A. and Dudley-Evans T. 1988. 'A Genre-Based Investigation of the Discussion Sections in Articles and Dissertations .' in *English for Specific Purposes Journal* Vol. 7/2: 113-121
Howarth, P. 1993. `A Phraseological Approach to Academic Writing', in G. Blue (ed.) *Language, Learning and Success: Studying Through English*, London: Macmillan,: 58-69.
Howarth P. 1996. *Phraseology in English Academic Writing. Some Implications for Language Learning and Dictionary Making*. Tübingen: Max Niemeyer Verlag.
Howarth P. 1998. 'The Phraseology of Learners' Academic Writing'. In A.P. Cowie (ed.) Pp161-186.
Huddleston R.D. 1971. *The Sentence in Written English. A Syntactic Study Based on an Analysis of Scientific Texts* Cambridge University Press.
Hunston S. 1993. 'Projecting A Sub-Culture: the Construction of Shared Worlds By Projecting Clauses in Two Registers.' in D. Graddol, L Thomson and M Byran (eds.) 1993. *Language and Culture* Clevedon: BAAL: 98-112
Hunston S. 1995. 'Ideology, Genre and Text in Systemic Linguistics.' Unpublished MS Presented at BAAL / CUP Genre Analysis Workshop, Sheffield July 1995.
Hunston, S. and Francis, G. 1998. `Verbs Observed: A Corpus-Driven Pedagogic Grammar', *Applied Linguistics*, 19/1, 45-72

Hymes D.H. 1971. *On Communicative Competence* Philadelphia: University of Pennsylvania Press

Ide N.M. 1993. 'A Statistical Measure of Theme and Structure.' *Computers and the Humanities* Vol. 13: 277-283

Inman B. 1978. 'Lexical Analysis of Scientific and Technical Prose.' in M.T. Trimble et al. (eds.) 1978: 242-56

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

Johansson S. 1982. 'Word Frequency and Text Type: Some Observations Based on the LOB Corpus of British English Texts.' in *Computers and the Humanities* Vol.19: 23-36

Johns T. and King P. 1993. *Data-Driven Learning Workshop* Presented at the B.A.L.E.A.P Meeting, University of Birmingham, March 22 1993

Johns T. and Scott M. 1994. *Microconcord Concordancing Programme.* Oxford University Press: Oxford.

Källgren G. 1988a. 'Automatic Indexing and Generating of Content Graphs from Unrestricted Text.' in Ö. Dahl and K. Fraurud (eds.): 147-160

Källgren G. 1988b. 'Automatic Abstracting of Content in Text.' in *Nordic Journal of Linguistics* Vol. 11 89-110

Kay P. and Fillmore C.J. 1999. 'Grammatical Construction and Linguistic Generalization: the *What's X Doing Y?* Construction.' in *Language* 75/1: 1-34.

Kaye G. 1990. 'A Corpus-Builder and Real-Time Concordance Browser for An IBM PC.' in J. Aarts and W. Meijs (eds) 1990: 137-161

Kennedy G. 1984. 'Preferred Ways of Saying Things with Implications for Language Teaching.' in J. Aarts and W. Meijs (eds) 1984: 335-373

Kennedy G. 1991. '*Between* and *Through :* the Company They Keep and the Functions They Serve.' in K. Aijmer and B. Altenberg (eds) 1991: 95-110

Kevles D. 1995. 'Pursuing the Unpopular: A History of Courage, Viruses and Cancer.' in R. Silvers (ed.) 1995 *Hidden Histories of Science*. New York: Granta,: 69-112.

Khurshid A. 1979. 'On Abstracts and Abstracting.' in *Annals of Library Science and Documentation* Vol. 26: 14-20

Kilgariff A. 1996. 'Comparing Frequencies across Corpora: Why Chi-Square Doesn't Work, and An Improved LOB-Brown Comparison.' *Proceedings of the Conference of the Association of Literary and Linguistic Computing-ACH 1996*, University of Bergen, June 25-29, 1996: 169-173

Kinay A.N., Muloshi L.P., Musakabantu M.R. and Swales J.M. 1983. 'Pre-Announcing Results in Article Introductions.' MS, Birmingham UK: Language Studies Unit, University of Aston

King R. 1976. 'A Comparison of the Readability of Abstracts with their Source Documents.' in *Journal of the American Society for Information Science* Vol. 27: 118-121

Kintsch W. 1993. 'Information Accretion and Reduction in Text Processing Inferences.' in *Discourse Processes* Vol. 16/1 193-202

Kintsch W. and Van Dijk T. 1978 'Towards a Model of Text Comprehension and Production' in *Psychology Review* Vol.85/5: 363-394

Kjellmer G. 1984. 'Some Thoughts on Collocational Distinctiveness.' in J. Aarts and W. Meijs (eds) 1984: 163-171

Kjellmer G. 1990. 'Patterns of Collocability.' in J.Aarts and W. Meijs (eds) 1990: 163-178

Knorr-Cetina K. D. (ed.). 1983. *Science Observed: Perspectives on the Social Study of Science* London: Sage

Koch C. 1991. 'On the Benefits of Interrelating Computer Science and the Humanities: the Case of Metaphor.' in *Computers and the Humanities* Vol. 25: 289-295

Kouřilova M. (Forthcoming) 'Interactive Functions of Language in Peer Reviews of Medical Papers Written By Non-Native Speakers of English' Unpublished MS.

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

Krishnamurthy R. 1987. 'The Process of Compilation.' in J.McH. Sinclair (ed.) 1987: 62-85

Kučera H. and Francis W. N. 1967. *Computational Analysis of Present Day American English* Providence: Brown University Press

Lackstrom S., Selinker L. and Trimble L. 1972. 'Grammar and Technical English.' in *English Teaching Forum* Sept-Oct.: 3-14

Lackstrom S., Selinker L. and Trimble L. (eds.) 1973. 'Technical Principles and Grammatical Choice.' in *TESOL Quarterly*¨Vol. 7: 127-136

Latour B. and Woolgar S. 1986. *Inside the Laboratory. the Construction of Scientific Facts* New York: Garland Press

Lakoff G. 1987. *Women, Fire and Dangerous Things. What Categories Reveal about the Mind.* University of Chicago Press: California

Leech G. 1991. 'The State of the Art in Corpus Linguistics.' in K. Aijmer and B. Altenberg 1991: 8-29

Leech G. 1992. 'Corpora and Theories of Linguistic Performance.' in J. Svartvik (ed) 1992a: 105-125

Leech G. and Fligelstone S. 1992. 'Computers and Corpus Linguistics.' in C. Butler (ed.): 115-140

Lehrberger J. 1982. 'Automatic Translation and the Concept of Sublanguage.' in R. Kittredge and J. Lehrberger (eds.) *Sublanguage: Studies of Language in Restricted Semantic Domains*, Berlin: Walter De Gruyter: Chapter 3.

Lemke J.L.1991. 'Text Production and Dynamic Text Semantics.' in E. Ventola (ed.) 1991: 23-37

Lemke J. L. 1998 'Multiplying Meaning. Visual and Verbal Semiotics in Scientific Text'. In J. R. Martin, R. Veel (eds) 1998 *Reading Science: Critical and Functional Perspectives on Discourses of Science*. London: Routledge. 87-113.

Lévi-Strauss C. 1962 *La Pensée Sauvage* Paris: Plon

Liddy E., Bonzi S., Katzer J., and Oddy E. 1987. 'A Study of Discourse Anaphora in Scientific Abstracts.' in *Journal of the American Society for Information Science* Vol. 38: 255-261

Linstromberg S. 1991. 'Metaphor and ESP: A Ghost in the Machine? *English for Specific Purposes* Vol. 10/3: 207-225

Ljung M. 1991. 'Swedish TEFL Meets Reality.' in S. Johansson and B. Stenström (eds.): 245-256

Love A. 1993. 'Lexico-Semantic Features of Geology Textbooks'. In *English for Specific Purposes* Vol.12/3: 197-218

Louw B. 1993. 'Irony in the Text Or Insincerity in the Writer? the Diagnostic Potential of Semantic Prosodies.' in Baker et al. (eds.) 1993: 157-176

Luhn H.P. 1968. 'Key-Word-in-Context Information Index for Technical Literature.' in C.K. Schultz (ed.) *H.P.Luhn: Pioneer of Information Sciences: Selected Works* New York: Spartan

Lundquist L. 1992. 'Some Considerations on the Relations Between Text Linguistics and the Study of Text for Specific Purposes.' in Schröder (ed.): 231-243

Lundquist L. 1989. 'Coherence in Scientific Text.' in W. Heydrich et al. (eds.): 122-149

Luzon-Marco, M.J. 1999. 'Corpus Analysis and Pragmatics: A Study of the Negative Structure *Fail to*' in *ITL-Review of Applied Linguistics*. 123/124: 37-55

Lyne A.A. 1975 'A Word-Frequency Count of French Business Correspondence.' in *IRAL* Vol. 13/2: 95-110

Lyne A. A. 1983. 'Word Frequency Counts: Their Particular Reference to the Description of Languages for Special Purposes and A Technique for Enhancing Their Usefulness'. In *Nottingham Linguistic Circular*. 12/2: 130-140.

McCarthy M. 1984. 'A New Look at Vocabulary in EFL'. In *Applied Linguistics* 5/1: 12-21.

McCarthy M. and Carter R. 1994. *Language As Discourse. Perspectives for Language Teaching New York: Longman*

McEnery T. and Wilson A. 1996. *Corpus Linguistics* Edinburgh University Press: Edinburgh

McKinlay J. 1983. 'An Analysis of the Discussion Section of Medical Journal Articles.' Unpublished MSc Thesis. ESP Collection, Language Studies Unit, Aston University

McKinney M. 1991. 'Experimenting on and Experimenting with: Polywater and Experimental Realism.' *British Journal of the Philosophy of Science* Vol. 42: 295-307

Makkai A. 1992. 'The Challenge of the Virtual Dictionary and the Future of Linguistics.' in *International Journal of Lexicography* Vol. 5/4: 252-269

Malcolm L. 1987. 'What Rules Govern Tense Usage in Scientific Articles?' in *English for Specific Purposes Journal* Vol. 6/1: 31-43

Malinowski B. 1923. 'The Problem of Meaning in Primitive Languages.' Supplement to C.K. Ogden and I.A.Richards (eds.) *the Meaning of Meaning* New York: Harcourt Brace Jovanovich

Martin J.R. 1989. *Ideation: the Company Words Keep* Cambridge: Cambridge University Press

Martin J.R. 1991. 'Nominalization in Science and Humanities: Distilling Knowledge and Scaffolding Text.' in E. Ventola (ed.) 1991: 307-337

Master P. 1987. 'Generic *the* in *Scientific American*'. In *English for Specific Purposes* Vol. 6/3: 165-186

Master P. 1991. 'Active Verbs with Inanimate Subjects in Scientific Prose.' in *English for Specific Purposes* Vol. 10/1: 15-33

Mauranen A. 1993. 'Theme and Prospection in Written Discourse.' Baker et al. (eds.) 1993: 95-114

Mel'čuk I. 1995. 'Phrasemes in Language and Phrasemes in Linguistics'. In Everaert et al. (eds.): 167-232.

Mel'čuk I. 1998. 'Collocations and Lexical Functions' in Cowie (ed).: 23-54.

Meijs W. (ed.). 1987. *Corpus Linguistics and Beyond* Amsterdam: Rodopi

Meijs W. 1992. 'Computers and Dictionaries' in C. Butler (ed.): 141-165

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

Miall D.S. 1992. 'Estimating Changes in Collocations of Key Words across A Large Text: A Case Study of Coleridge's Notebooks.' in *Computers and the Humanities* Vol. 26: 1-12

Moon R. E. 1987. 'The Analysis of Meaning.' in J. McH. Sinclair (ed.) 1987: 86-103.

Moon R. E. 1992. 'There Is Reason in the Roasting of Eggs. A Comparison of Fixed Expressions in Native Speaker Dictionaries.' in *Euralex '92 Proceedings* Oxford University Press: 493-502

Moon R.E. 1994. 'The Analysis of Fixed Expressions in Text'. In M. Coulthard (ed). Pp117-135.

Moon, R.E. 1998a. *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. (Oxford Studies in Lexicography and Lexicology) Oxford: Oxford University Press.

Moon R.E. 1998b. 'Frequencies and Forms of Phrasal Lexemes in English'. In A. P. Cowie (ed).: 79-100.
Moskovitch G.M. and Caplan A. 1979. 'Distributive Statistical Techniques in Linguistic and Literary Research. " in D.E.Ager, F.E. Knowles and J. Smith (eds.): 245-263
Muller C. 1968. *Essai de Statistique Léxicale* Paris: Librairie Klincksieck
Muller C. 1977. *Principes et Méthodes de Statistique Léxicale* Paris: Hachette Université
Myers G. 1989. 'The Pragmatics of Politeness in Scientific Articles.' in *Applied Linguistics* Vol. 10 / 1: 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.' in *Discourse Processes* Vol. 14/1: 1-26
Myers G. 1992. 'Textbooks and the Sociology of Scientific Knowledge.' in *English for Specific Purposes* Vol. 11: 3-17
Nattinger J.R. and DeCarrico 1992. *Lexical Phrases and Language Teaching* Oxford: Oxford University Press
Nattinger J.R. and DeCarrico 1989. 'Lexical Acts and Teaching Conversation.' in *Vocabulary Acquisition: AILA Review 6*: 118-139
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 Ventola (ed) 1991: 369-384
Nystrand M. 1982. *What Writers Know. The Language, Process and Structure of Written Discourse* New York: Academic Press
Nystrand M. 1986. *The Structure of Written Communication: Studies in Reciprocity Between Writers and Readers* Orlando Fl.: Academic Press
Oakes M. 1996. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.
Oppenheim R. 1988. 'The Mathematical Analysis of Style: A Correlation-Based Approach.' in *Computers and the Humanities* Vo.22: 241-253
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
Papegaaij and Schubert R. 1988. *A Corpus-Based Bilingual Knowledge Bank for Distributed Language Translation* DLT Publications Amsterdam.
Pavel S. 1993a. 'Neology and Phraseology as Terminology-in-the-Making.' in H.B. Sonneveld and K.L.Loening (eds.) 1993: 21-34
Pavel S. 1993b. 'La Phraséologie en Langue de Spécialité. Méthodologie de Consignation dans les Vocabulaires Terminologiques.' Unpublished MS, Secrétariat d'État du Canada: Direction de la Terminologie et des Services Linguistiques.
Pavel S. and Boileau P. 1994. *Systèmes Dynamiques et Imagerie Fractale. Vocabulaire Français-Anglais.* Secrétariat d'État Du Canada: Direction De La Terminologie Et Des Services Linguistiques. Canada
Pawley A. and Syder F.H. 1983. 'Two Puzzles for Linguistic Theory: Naturelike Selection and Naturelike Fluency.' in Richards and Schmidt (eds.) 1985 *Language and Communication* London: Longman: 191-226.
Pearson J. 1998. *Terms in Context*. Amsterdam: John Benjamins.
Pettinari C. 1982. 'The Function of A Grammatical Alteration in 14 Surgical Reports.' in W. Frawley (ed.) 1982: 145-183.

Phillips M. 1985. *Aspects of Text Structure: An Investigation of the Lexical Organization of Text* Amsterdam: Elsevier NHL Series

Phillips M. 1989 *Lexical Structure of Text* Discourse Analysis Monograph No. 12, Birmingham: English Language Research, University of Birmingham

Picht H. and Draskau J. 1985. *Terminology: An Introduction* Surrey University Department of Linguistic and International Studies Monographs.

Popiel S. and K. McRae. 1988. 'The Figurative and Literal Senses of Idioms, Or All Idioms Are Not Used Equally.' in *Journal of Psycholinguistic Research* 17/6: 475-487.

Potter R. G. 1991. " Statistical Analyses of Literature: A Retrospective on *Chum*: 1966-1990' in *Computers and the Humanities* Vol. 25: 401-429

Propp V. 1968*1928. *The Morphology of the Folktale* University of Texas Press

Quirk R. 1995. *Grammatical and Lexical Variance in English* London: Longman.

Quirk R., Greenbaum S., Leech G. and Svartvik J. 1985. *A Comprehensive Grammar of the English Language* London: Longman.

Raya F. 1986. 'Writing Abstracts for Free-Text Searching.' in *Journal of Documentation* Vol. 42: 11-21

Reder L.M. and Anderson J.R. 1980. 'A Comparison of Texts and their Summaries; Memorial Consequences.' in *Journal of Verbal Learning and Verbal Behaviour* Vol. 19: 121-134

Renouf A. 1987a. 'Lexical Resolution.. In W. Meijs (ed.) 1987

Renouf A. 1987b. 'Corpus Development.' in J. McH. Sinclair (ed) 1987: 1-41

Renouf A. 1991. 'Coding Metalanguage: Issues Raised in the Creation and Processing of Specialised Corpora.' in S. Johansson and B. Strenström (eds.): 198-206

Renouf A. 1998. *Explorations in Corpus Linguistics*. (Language and Computing 23). Rodopi: Amsterdam.

Renouf A. and Sinclair J. McH. 1991. 'Collocational Frameworks in English.' in K. Aijmer and B. Altenberg 1991: 128-144

Richards J.C. and Schmidt R. (eds.) 1983. *Language and Communication* London: Longman

Ringle M. 1982. 'Artificial Intelligence and Semantic Theory.' in T.W. Simon and R.J. Scholes (eds.) *Language, Mind and Brain* London: Erlbaum

Roe P. J. 1993a. 'ASTEC: Users' Guide to the Aston Corpus of Scientific and Technical English.' Internal Report, Language Studies Unit: Aston University

Roe P. J. 1993b. 'Software Specification for ATA (Aston Text Analyser).' Internal Report, Language Studies Unit: Aston University

Rundell M. and Stock P. 1992. 'The Corpus Revolution.' *English Today* April-October 1992

Sager J.C. 1990. *A Practical Course in Terminology Processing* Amsterdam: John Benjamins

Sager J.C., Dunkworth D. and P.F. McDonald 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.' in *English for Specific Purposes* Vol. 11/2: 93-114

Salager-Meyer F. 1990a. 'Metaphor in Medical English Prose: A Comparative Study with French and Spanish.' in *English for Specific Purposes* Vol.9: 145-159

Salager-Meyer F. 1990b. 'Discoursal Flaws in Medical English Abstracts' in *Text* Vol. 10/4: 365-384

Sampson G. and Haigh R. 1988. 'Why Are Long Sentences Longer Than Short Ones?' in M. Kytö et al. (eds.): 207-219

Sastri M. 1968. 'Prepositions in Chemical Abstracts.' in *Linguistics* Vol. 38: 23-28

Saussure de F. 1916. *Cours De Linguistique Générale.* Paris: Payot.

Saville-Troike M. 1982. *The Ethnography of Communication* Oxford: Basil Blackwell
SCIENCE CITATION INDEX 1993. *Journal Citation Reports* Institute for Scientific Information: Philadelphia
Schank R.C. and Abelson R.P. 1977. *Scripts, Plans, Goals and Understanding. An Inquiry Into Human Knowledge Structures* New Jersey: Lawrence Erlbaum
Schiffrin D. 1990. 'Between Text and Context: Deixis, Anaphora and the Meaning of *Then*' in *Text* 10/3: 245-270
Schubert K. 1986. *Distributed Language Translation* Amsterdam: Elsevier Science
Scott W.A.H. 1991. *Chemistry* Glasgow: Harper Collins
Scott M. 1993. 'Lexical Tools for Genre Analysis for Computers.' Unpublished MS Presented at the BAAL Annual Meeting 14-16 Sept. 1993
Searle J.P. 1969. *Speech Acts* London: Oxford University Press
Sharp B. 1989. 'Elaboration and Testing of New Methodologies for Abstracting' Unpublished Ph.D Thesis, Modern Languages Department, Aston University
Sherrard B. 1989. 'Teaching Students to Summarize: Applying Textlinguistics.' in *System* Vol. 17/1: 1-11
Sinclair J. McH. 1980. 'Some Implications of Discourse Analysis for ESP Methodology.' in *Applied Linguistics* 1/3: 253-261
Sinclair J. McH. 1981. 'Planes of Discourse.' MS, English Department of the University of Birmingham, Presented in S.N.A. Rizvil (ed.) 1983 *the Two-Fold Voice: Essays in Honour of Ramesh Mohan* at the University of Salzburg
Sinclair J.McH. 1984. 'Naturalness in Language.' in J. Aarts and W. Meijs (eds.) 1984: 203-210
Sinclair J. McH. (ed.) 1987a. *Looking Up: An Account of the Collins COBUILD Project* London: Collins ELT
Sinclair J. McH. 1987b. 'Grammar in the Dictionary': 104-115 and 'The Notion of Evidence.': 130-159 in J. McH. Sinclair (ed.) 1987a.
Sinclair J. McH. 1987c. 'Collocation: A Progress Report.' in R. Steele and T. Threadgold (eds.) *Language Topics: Essays in Honour of Michael Halliday*. 1987: Amsterdam: John Benjamins: 319-331
Sinclair J. McH. 1988. 'Compressed English.' in M. Ghadessy (ed.) 1988: 130-136
Sinclair J. McH. 1991. *Corpus, Concordance, Collocation* Oxford: Oxford University Press
Sinclair J. McH. 1992. 'The Automatic Analysis of Corpora.' in J. Svartvik (ed.) 1992: 379-397
Sinclair J. McH. 1993a. 'Text Corpora: Lexicographer's Needs.' in *Zeitschrift für Anglistik und Amerikanistik* Vol. XLI: 1/1: 5-13
Sinclair J. McH 1993b. 'Posturing in Discourse.' Keynote Speech Presented at the 5th International Systemic Workshop on Corpus-Based Studies, Universidad Complutense De Madrid, 26-29 July 1993
Sinclair J. McH 1993c. 'The Bank of English: A British and International Corpus of English.' in *Zeitschrift Für Anglistik Und Amerikanistik* Vol. XLI 2/2: 166-167
Sinclair J. McH. 1993d. 'Written Discourse Structure.' in J.McH Sinclair et al. (eds.) 1993: 6-31
Sinclair J. McH. 1994. 'Trust the Text'. In M. Coulthard (ed.) London: Routledge. Pp12-25.
Sinclair J. McH., Hoelter M., and Peters C. (eds.) 1995. *the Languages of Definition: the Formalisation of Dictionary Definitions for Natural Language Processing*, Luxemburg: Office for Official Publications of the European Committees.
Sinclair J., McH. Hoey M., and Fox G. (eds.) 1993. *Techniques of Description: Spoken and Written Discourse* London: Routledge

Sinclair J. McH., Jones S. and Daley R. 1969. *English Lexical Studies*. UB Report for the Office of Science and Technology Information.

Smadja F. 1993. 'Retrieving Collocations from Text: Xtract.' in *Computational Linguistics* Vol19/1: 143-177

Smadja F. 1996. 'Translating Collocations for Bilingual Lexicons: A Statistical Approach'. In *Computational Linguistics* 22/1 Pp1-38.

Sonneveld H.B. and Loening K.L. (eds.) 1993. *Terminology. Applications in Interdisciplinary Communication*. John Benjamins: Amsterdam

Souter C. 1990. 'Systemic-Functional Grammars and Corpora.' in Aarts and Meijs (eds.) 1990: 179-211

Sparck-Jones K. 1971. *Automated Keyword Classification for Information Retrieval* London: Butterwoth

Stubbs M. 1982. 'Written Language and Society: Some Particular Cases and General Observations.' in M. Nystrand (ed.) 1992: 31-55

Stubbs M. 1987. 'An Educational Theory of (Written) Language.' in T. Bloor and J. Norrish (eds.) *BAAL 2: Papers from the Annual Meeting of the British Association for Applied Linguistics* London, CILT: 3-38

Stubbs M. 1993. 'British Traditions in Text Analysis. from Firth to Sinclair.' in M. Baker et al. (eds.) 1993 1-33

Stubbs M. 1994. 'Grammar, Text and Ideology: Computer-Assisted Methods in the Linguistics of Representation'. In *Applied Linguistics* Vol.15/2: 201-223

Stubbs M. 1996. *Text and Corpus Analysis* Routledge: London.

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

Svartvik J. 1992b. 'Corpus Linguistics Comes of Age.' : 7-13 in J. Svartvik 1992a

Svartvik J. 1993. 'Lexis in English Language Corpora.' in *Zeitschrift Für Anglistik Und Amerikanistik* Vol. XLI: 1/1: 13-31

Swales J. 1981a. *Aspects of Article Introductions* Aston ESP Research Report No.1, Language Studies Unit: Aston University

Swales J. 1981b. 'Definitions in Science and Law: A Case for Subject Specific ESP Materials.' in *Fachsprache* Vol. 81/3: 106-112

Swales J. 1981c. 'The Function of One Type of Particle in A Chemistry Textbook.' in Selinker et al. (eds.): 40-52

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

Swales J. 1998. *Other Floors, Other Voices*. *A Textography of A Small University Building*. Mahwah, N.J. Lawrence Erlbaum.

Swales J. and Najjar H. 1987. 'The Writing of Research Article Introductions.' in *Written Communication* Vol. 4: 175-192

Tarasova T. 1993. 'Non-Verbal Elements in Scientific Text.' Unpublished Ph.D. Thesis, Language Studies Unit, Aston University.

Thomas P. 1993. 'Choosing Headwords from LSP Collocations for Entry Into A Terminology Data Bank (Term Bank).' in Sonneveld H.B. and Loening K.L. (eds.) 1993: 46-68.

Thomas H. and Waxman J. 1995. 'Oncogenes and Cancer.' in J. Waxman and K. Sikera (eds.) *the Molecular Biology of Cancer*: 1-17.

Thompson G. and Yiyun Y. 1991. 'Evaluation in the Reporting Verbs Used in Academic Papers.' in *Applied Linguistics* Vol. 12/4: 365-382

Traugott E. and Heine H. 1991. *Approaches to Grammaticalisation*. Vol. II. Amsterdam: John Benjamins.

Ure J. 1971. 'Lexical Density and Register Differentiation.' in G. E. Prerren and J.L.M. Trim (eds.) *Applications of Linguistics* Cambridge: Cambridge University Press

Van Der Wouden T. 1997. *Negative Contexts. Collocation, Polarity and Multiple Negation*. Routledge: London.

Van Dijk T. 1979. *Macrostructures: An Interdisciplinary Study of Global Structures in Discourse* Hillsdale New Jersey: Lawrence Erlbaum

Van Dijk T. and Kintsch W. 1983. *Strategies of Discourse Comprehension* New York; Academic Press

Van Dijk T. and Kintsch W. 1978. 'Cognitive Psychology and Discourse: Recalling and Summarizing Stories. " in W. Dressler (ed.) *Current Trends in Textlinguistics.* Berlin: De Gruyter.

Van Halteren H. 1994. 'Syntactic Databases in the Classroom.' in Wilson and McEnery (eds.): 17-28

Van Roey J. 1990. *French-English Contrastive Lexicology: An Introduction*. Louvain-La-Neuve: Peeters.

Varttala T. 1999. 'Remarks on the Communicative Function of Hedges in Popular Scientific and Specialist Research Articles.' in *English for Specific Purposes*. 18/2: 177-200.

Ventola E. (ed.) 1991. *Functional and Systemic Linguistics: Approaches and Uses* Den Haag: Mouton De Gruyter

Ventola E. and Mauranen A. 1991. 'Non-Native Writing and Native Revising of Scientific Articles.' in E. Ventola (ed.): 457-492

Verschueren J. 1999. *Understanding Pragmatics*. London: Arnold.

Vidalenc J-L. 1997. 'Quelques remarques sur l'emploi de la métaphore comme outil de dénomination dans un corpus d'histoire des sciences.' in Boisson C. and Thoiron P. (eds.) 1997. *La Dénomination*. Paris: Presses Universitaires De Lyon.: 1-11.

Vossen P., den Broeder M. and Meijs W. 1986. 'The LINKS Project: Building A Semantic Database for Linguistic Applications.. In Aarts and Meijs (eds.) 1986: 277-293

Weil B.H., Zarember I. and Owen H. 1963. 'Technical Abstracting Fundamentals. Part II. Writing Principles and Practices.' in *Journal of Chemical Documentation* Vol. 3/1: 125-132

West G.K. 1980. '*That*-Nominal Constructions in Traditional Rhetorical Divisions of Scientific Research Papers.' in *TESOL Quarterly* Vol. 14: 483-489

Wikberg K. 1990. 'Topic, Theme and Hierarchial Structure in Procedural Discourse.' in J. Aarts and W. Meijs (eds.) 1990: 281-254

Wilbur W.J. and Sirotkin K. 1992. 'The Automatic Identification of Stop Words.' in *Journal of Information Science* Vol. 18/1: 45-55

Williams I. 1996. *'Ifs* and *Buts*. Impact Factors of Journals may Affect Decisions on Resource Allocation'. In *Chemistry in Britain*, February 1996: 31-33

Williams I. A. 1996. 'A Contextual Study of Lexical Verbs in Two Types of Medical Research Article.' in *English for Specific Purposes*. Vol 15/3: 175-198.

Willis D. 1990. *the Lexical Syllabus* London: Collins ELT

Willis D. 1993. 'Grammar and Lexis: Some Pedagogical Implications.' in Sinclair et al. (eds.) 1993: 83-93

Wilson A. and McEnery T. (eds) 1994. *Corpora in Language Education and Research: A Selection of Papers from Talc94.* UCREL Technical Papers 4., Lancaster University.

Wingard P. 1981. 'Some Verb Forms and Functions in Six Medical Texts.' in L. Selinker, E. Tarone and V. Hanzeli (eds.) *English for Academic and Technical Purposes: Studies in Honour of Louis Trimble*: 53-64

Winter E. 1977. 'A Clause Relational Approach to English Texts: A Study of Some Predictive Lexical Items in Written Discourse'. In *Instructional Science*. Vol. 6/1:1-92.

Winter E. 1996. 'Metalanguage Nouns of Clause Relations' Unpublished paper presented at *Corpus Research: Sharing Interpretations*. CELS, University of Birmingham, 20th Sept. 1996.

Wittgenstein L. 1957. *Philosophical Investigations* Oxford: Blackwell

Wood P. 1982. 'An Examination of the Rhetorical Structures of Authentic Chemistry Texts.' in *Applied Linguistics* Vol. 3: 121-143

Yang H.Z. 1986. 'A New Technique for Identifying Scientific and Technical Terms and Describing Scientific Texts.' in *Literary and Linguistic Computing* Vol.1/2: 93-103

Youmans G. 1991. 'A New Tool for Discourse Analysis: the Vocabulary Management Profile.' in *Language* Vol. 67/4: 763-789

Yorio C. 1980. 'Conventionalized Language Forms and the Development of Communicative Competence.' *TESOL Quarterly*. Vol. 14/4: 433-422.

Yorio C. 1989. 'Idiomaticity as an Indicator of Second Language Proficiency'. In K. Hyltenstam and L. Obler (eds.) *Bilingualism across the Life-Span.* Cambridge: Cambridge University Press Pp55-72.

Zambrano S. 1987. 'A Comparison of the Linguistic Features and Discourse Structure of Abstracts and Conclusions' Unpublished M.Sc. Thesis, Language Studies Unit: Aston University.