SCIENTIFIC INNOVATION AND THE PHRASEOLOGY OF RHETORIC. POSTURE, REFORMULATION AND COLLOCATION IN CANCER RESEARCH ARTICLES.

CHRIS GLEDHILL

Doctor of Philosophy

THE UNIVERSITY OF ASTON IN BIRMINGHAM

July 1995

THE UNIVERSITY OF ASTON IN BIRMINGHAM

SCIENTIFIC INNOVATION AND THE PHRASEOLOGY OF RHETORIC.
POSTURE, REFORMULATION AND COLLOCATION IN CANCER RESEARCH ARTICLES.

SUMMARY

This thesis aims to describe how language is used in hard science, how scientists create new science in their writing and how language functions in extremely specialised circumstances. The thesis describes the working context of cancer research articles at Aston University's Pharmaceutical sciences department. The thesis attempts to integrate the ethnographic approach of genre analysis (Swales 1990) with the large scale analysis of phraseology in the field of corpus linguistics Sinclair (1987a).

One hypothesis is that new science is actually enacted in research articles by a process of reformulating concepts within the text. To test this, the scientific claims of a sample of ten texts are analysed in terms of reformulation of grammatical metaphor, discourse signalling and posture (Halliday 1985, Sinclair 1981). A second hypothesis is that new science is founded on a system of preferred expressions, and that collocation is a fundamental mechanism that allows for new formulations to take place throughout the text. A corpus analysis of 150 cancer research articles is undertaken to characterise the phraseology of grammatical items in research articles and in the various rhetorical sections of research articles namely Titles, Abstracts, Introductions, Methods, Results and Discussion sections.

The thesis finds that research articles use language to create new science by reformulating data as research models and by altering the established patterns of phraseology. Collocation is seen to vary systematically in rhetorical sections, and the concept of phraseology is postulated as a preferred way of expressing a delimited set of semantic and communicative roles. The thesis argues that science should not be seen as a body of facts transmitted via language, but as a special linguistic construct, mediated by the mechanisms of textual reformulation and phraseological innovation.

Key words:

cancer research
collocation
corpus linguistics
reformulation
English for specific purposes

Chris Gledhill

PhD
1995

Ceci est pour Céline.

ACKNOWLEDGEMENTS

I am especially grateful to the following members of staff from the Pharmaceutical Sciences Departments at Aston and Birmingham Universities for their participation in my survey and for allowing me to use their publications: Dominique Armspach, William Fraser, Sally Freeman, John Gardiner, Andy Genscher, Helen Mulligan, William Irwin, Philip Lambert, Richard Lewis, Peter R. Lowe, David Poyner, Michael Tisdale, Yaruko Wang and Richard Wheelhouse. I would especially like to thank Professor Tisdale for his advice on cancer research and Bill Fraser and Sally Freeman for helping me with access to the electronic indexes and the department.

I would also like to extend my thanks to Professors Frank Knowles for his expert advice on corpus linguistics, Mike Hoey for his ideas on lexis, John Sinclair for his insights into discourse and Professors Denis Ager, John Gaffney and Nigel Reeves who were instrumental in their support for me at Aston. Mike Scott at Liverpool also deserves my special thanks, not only for allowing me to use his suite of programs but also for his constant encouragement and interest since the project began.

I would also like to thank the following colleagues and friends for their ideas, expertise and enthusiasm at various stages of the thesis: Meriel Bloor, Chris Butler, Beverley Derewianka, Julian Edge, Gill Francis, Tim Gibson, Robin Goodfellow, Susan Hunston, Agnes Kukulska-Hulme, Paul Meara, Lorna Milne, Peter Roe, Christina Schäffner, Patricia Thomas and Jane Willis. I also owe a great deal to my friends from esperantujo, Don Lord and Dafne Lister, who introduced me to linguistics. I am also indebted to the British Academy in awarding me its valuable financial support and to my Grandfather when it ran out. I would also like to mention here my late uncle Steven Gibson who influenced my thinking in many ways.

Most of all however, I would like to extend my warmest thanks of all to Tom Bloor, my teacher and supervisor. There are many of his ex-students teaching language throughout the world who have benefited from his thorough and genial approach to language.

LIST OF CONTENTS

PART I INTRODUCTION AND LITERATURE REVIEW **Page**

CHAPTER FOUR: REFORMULATION IN SCIENTIFIC DISCOURSE.

CHAPTER FIVE: PHRASEOLOGY AND CORPUS LINGUISTICS.

PART II SURVEY AND DATA COLLECTION

CHAPTER SIX: RESEARCH SYNTHESIS.

CHAPTER SEVEN: RESEARCH CONTEXT

CHAPTER EIGHT: THE CORPUS

PART III DATA ANALYSIS

CHAPTER NINE: REFORMULATION IN PHARMACEUTICAL SCIENCES ARTICLES.

CHAPTER TEN: POSTURE AND DISCOURSE SIGNALLING.

CHAPTER ELEVEN: PHRASEOLOGY IN THE PHARMACEUTICAL SCIENCES CORPUS.

LIST OF TABLES AND FIGURES

PART I: INTRODUCTION AND LITERATURE REVIEW

This thesis aims to contribute to a description of scientific discourse. The topic is limited to discussion of two specific types of scientific text, the abstract and article in refereed journals in the fields of cancer research and pharmacology. In the first section of this introduction, the motivation of the thesis is set out, followed by three sections justifying the topic of research in terms of scientific discourse and linguistic interest. On the basis of this, three preliminary research questions and their corresponding research organisation are formulated, and then addressed by a review of current work on the subject in Chapter 2.

CHAPTER ONE: INTRODUCTION

1.0 Research interests.

> Les mots sont modelés sur des objets à notre échelle. Ils ont acquis leur efficacité en s'adaptant à des phénomènes ou à des événements de notre monde quotidien. Aussi, quand on aborde des réalités à une autre échelle, les mots deviennent facilement des obstacles. (Reeves 1981:68)

This thesis is motivated by an interest in how language changes in specific circumstances, how it interacts with our conceptions of the world and how it may form them. In the sciences, as the Canadian astrophysicist Hubert Reeves says above, language is not perceived as a constructive way of representing knowledge, indeed many scientists and some linguists see knowledge as a body of non-linguistic expressions (De Beaugrande and Dressler 1981:85, Escarpit 1976, Van Dijk and Kintsch 1973). Indeed, scientific knowledge is thought to be at its most refined when it is expressed in its own abstract symbolic system, non-verbally in algebraic and logical formulae, atomic structural models, tree-diagrams and so on (Auger 1975, Cremmins 1975). For example in chemistry, structural diagrams are capable of expressing entire chemical processes (themselves termed 'stories') as well as the chemical compounds themselves. But if we look at research articles in the journals Fractal Geometry or Celestial Mechanics, their abstract symbolism is still accompanied by tiny outcrops of language, often integrated with the rest by way of punctuation and text format. These esoteric texts are surrounded by language: the natural language summary, the title, contents, copyright statements, introductions and covers of periodicals and books are all peripheral elements of context that Lane (1992) calls the "péritexte".

Our preoccupation with science takes place in the context of a broad philosophical consensus on the relation between science and language, namely 'constructivism', a paradigm that pervades a broad spectrum of disciplines concerned with science writing. For example, at the intersecting edge of science and linguistics, some terminologists have proposed the constructivist view that language augments scientific reality. For example, Sager (1990, and others) have consistently argued that this relationship is a metalinguistic one: that language frames the logical acceptability of theorems. Similarly, linguists writing about the language of science (such as Swales 1990, Myers 1990, Halliday and Martin 1993) have based their observations on the findings of social anthropologists, that science is a social-construct negotiated by the conflicting forces of discourse communities (Latour and Woolgar 1972, Knorr-Cetina 1983, Kaplan and Grabe 1992). Philosophers of science, such as McKinney (1991) equally emphasise the discoursal nature of science and mention the famous case of polywater, the wonder substance which inspired a new but short-lived chemical paradigm. In an interesting mirror-image of research activity, scientists themselves have even pondered on the evolutionary aspects of cultural transmission (Cavalli-Sforza and Felman 1981, Dawkins 1986). The constructivist paradigm emerges as a powerful model. It emphasises the role of language in scientific research, it claims that there is a causal relation between discourse and scientific fact, and centrally to our thesis it suggests that there are mechanisms of language which enable not only the transmission but also the creation of scientific knowledge.

The underlying motivation of this thesis is to explore what happens when language is used in an extremely specialised context. This kind of delimited language use is the subject of English for Specific Purposes (ESP), a field concerned with elaborating teaching methods that take account of the relationship between language and specialist subjects. The term 'specific' in ESP denotes the linguistic activities of a group of people with some joint purpose (Swales 1990:9). Consequently much of the work of ESP has been concerned with the language needs of professional groups, especially in science and technology; areas where English is still a dominant language as Sager et al. (1980:xviii) and Swales (1990:87) point out. The key problem for ESP is to find out what exactly constitutes the linguistic activities of these groups, and to characterise them for rapid transmission to other users.

The primary practical aim of this thesis is therefore to provide a characterisation of the specific linguistic practices of a professional group. In particular, this group consists of pharmacologists and cancer researchers at Aston University's pharmaceutical sciences

department. There are a number of practical reasons for this choice:

-Cancer research has received little attention from linguists.
-Cancer research is possibly one of world's biggest medical research activities, served by a large selection of the most prestigious journals.
-Cancer involves a broad sweep of specialisms (drug synthesis, genetics, patient care) that are all integrated into the global aims of the researchers at Aston.
- The field involves a high degree of abstract pharmaceutical knowledge that has a complex non-verbal system of representation with articles that are written in a very highly refined English.
-The cancer research department at Aston is an important research centre for the UK serving the National Cancer Institute and it has an above-average output of research with a number of high profile breakthroughs reported in the media.

Before presenting current research on this area, we set the general context of research on science writing.

1.1 Research orientation.

Linguistics offers two apparently distinct views of the relation between science and language:

1) The study of the language needs of specialists as a distinct social group, as elaborated by Swales (1981b,1990) in the field of ESP.
2) The study of the relationships between the specialist subject and language, in the fields of terminology and Languages for Special Purposes (Sager et al. 1980, Sager 1991).

Each approach has a different emphasis on language, although they intersect on several points and derive their linguistic standpoints from theories of language in use. The Firthian-Hallidayan school of language is the model we adopt here. This approach interprets language as a functional construct of society and as a construct (as well as 'constructor') of human knowledge, clearly a useful model that addresses the concerns of the ESP researcher and the terminologist. J.R.Firth is seen as the originator of systemics, and the relevance of his views to the study of language practice are evident:

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

The semiotic system is thus considered to operate within a social framework of communication, and it informed Halliday's conception of *discourse:*

> 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 recognize discourse... (Halliday 1977:165).

For these reasons, the systemic school has been particularly fruitful in providing a theoretical and methodological background to the study of scientific discourse (Ventola 1991, Mauranen 1991, Halliday and Martin 1991 *inter alia*).

But linguistic practice in science has not been a preoccupation of mainstream linguistics. The generative syntax of Chomsky (1957) and the formal discrete grammars which emerged from or in opposition to Chomsky's original models (c.f. Lyons et al 1987) are essentially concerned with the specification of rules to describe potential expression in language. The formal logical approach influenced the early stages of artificial intelligence (Charniak and Wilks 1976). The idea of semantic or cognitive primitives in syntactic or textual models of language has followed parallel lines in case grammar (Fillmore 1968), goal-orientated textgrammars (Schank and Abelson 1977) and proposition-oriented textgrammars (Van Dijk and Kintsch 1973). While transformational models and textgrammars have been used to describe certain scientific styles of writing (Gopnik 1972, Hutchins 1977, 1978) and syntactic approaches provide the theoretical basis of description for computational analysis of some large corpora (Aijmer and Altenberg 1991, Crystal 1991) they have had little application to a description of language practice. Instead the principle of linguistic competence has fuelled models of language acquisition rather than applied linguistics (Cook 1988, Gazdar 1987:123). A typical criticism of the formal grammar is that language is chiefly seen to play the role of an encoding and decoding device for information, a critical factor being the formalists' insistence on the distinction between formal and functional.

In opposition to this, descriptions of language that attempt a performance-oriented or communicative approach have paralleled relativist and hermeneutic philosophy (Wittgenstein 1957, Heidegger 1966, Gadamer 1976) in rejecting the idea that language can be described metalinguistically or in terms of truth values. Instead, scientific truth cannot be anything but 'rooted' (Heidegger's term) in its culture. The 'natural language philosophers' (Austin 1959, Searle 1969 and Grice 1975) also came to reject truth values, and instead

established a framework for the field of discourse analysis. They saw meaning as conventionalised in language rather than algorithmically encoded in it. A similar view of language use was championed by Lévi-Strauss (1962) and Barthes (1965) in the semiotic construction of social mythology.

Semiotics emerged from De Saussure's (1916) view of the creation of meaning as the result of a structural code rather than as the relationship with external truth or 'reality'. If the Firthian approach differs from this, it is in the idea that language is considered as the place not only for the passing on of information, but also as the medium for the binding of social relations. Firth claims that language in its very substance reflects the various levels of physical and communicative meaning of the context of situation (Malinowski 1923). Firth's principle of 'modes of meaning' (1957:190) is in this thesis taken to indicate that instead of linguistic forms being equally meaningful in the setting of a particular text (such as *nominalisation* or *rhetorical move*), the analysis should proceed from the basis that the text is a unique event utilising linguistic forms in a novel way (no matter how slight) and functioning within the constraints of a particular set of practices (despite resemblances to others). The implications of Firth's ideas are taken up later in our discussion of Halliday.

The relevance of Firth to the language of science is that the scientific text can be seen as a barometer of the social and professional context from which it emerges, changing as the social variables, textual conventions or topic change. In choosing the research article as the privileged place of scientific discourse (the reasons for which are set out in the next section), we need to attempt an understanding of the working practices of scientific research, including the world outside the laboratory: attending conferences, submitting articles to refereed journals, keeping up with the specialist literature and so on. But these variables and effects should not imply that an instance of language use is completely predictable, that language works in a mechanistic and reactive way. We argue below that this view is incompatible with the importance of written texts in scientific communities. An atomistic 'information' view of language, posited by information scientists such as Escarpit (1976) would be contradicted by any evidence to suggest that new scientific research is not just 'passed on' via language, but embodied in language. Rather, Firth proposes that not only do the social 'external' factors involved in the production of scientific texts have to be taken into consideration, something of the symbolic status of the text as part of the linguistic (semiotic) meaning system has to be realised and contrasted to other systems before a characterisation may take place. We make two fundamental assumptions on the basis of Firth's approach:

1- That the descriptive analysis of language, and in particular that of grammar, has to be rooted in context, here viewed as the discourse of the participants as opposed to intuitive *a priori* grammatical categories.
2- That grammar cannot be treated in isolation from the subject matter and consequently from the system of wording, or the lexico-grammar (Halliday 1991a).

It is difficult to see, perhaps, how a description may 'emerge' from the data without previous categories at hand. Guba and Lincoln (1992), writing in the field of educational evaluation, have set out a statement of good practice for the social sciences which allows the analyst ways of formulating such a description. The methodological issues implied by this are discussed later in Chapter 6. Here it suffices to say that Swales' (1990) analysis shares Guba and Lincoln's, as well as Firth's, concern for the holistic nature of social systems.

The linguists' terminology of 'semiotic system of science', 'scientific discourse community' and 'research article genre' form a preliminary conceptual framework for research into the language of cancer research. There are problems in their application, and this is discussed in Chapter 2. But these rough concepts transform our interest in specific language use into a research orientation concerned with discourse. Now that we have announced some very rough assumptions, three general questions can be asked in order to frame the research topic in more detail:

Preliminary question: What role do research articles have in scientific activity?

Specific question 1: What role does the research article have in the creation of new science?

Specific question 2: How does language function in extremely specialised contexts?

These questions are reformulated as Hypotheses in Part II, and below we set out preliminary answers in terms of the information value of language in science (general question), and the interaction of scientists and text (specific questions 1 and 2). The general question acts as a theoretical backdrop of the thesis: the specific questions are later addressed by linguistic analysis of reformulation and phraseology.

1.2 The information value of research articles in science

> Information has succeeded raw materials and energy as the primary commodity. (from Bell's The Post-Industrial Society cit. Auger 1989:iv)

The role of language in the scientific community has been equated with the management of information. As Maizell et al. (1971) note, the processes of information production, control and retrieval adapt to society's increasing dependence on information and to the increasing technological advances in information access. As change in society is reflected in society's shifting goals and needs, especially in the "mission-oriented fields of knowledge" as Auger puts it (1989:vi), processes of information access change form, or take on new roles. As new text forms appear, old forms change or disappear, as Atkinson (1992) has demonstrated in his analysis of the *Edinburgh Medical Journal* from individual personal report to public declaration. The diversification of texts is mirrored by the increased specialisation in fields of research resulting in a kind of discourse evolution where one field is seen to develop or expand while others split and diversify (Sager et al. 1980:xviii). One of the results of this is an increasing array of competing types of message form, including internet bulletins, computer accessible indexes, automatic search indexes, and interactive self-updating databases (Jennings 1990).

In theory, the time spent by researchers on reading articles and keeping up with their fields is bound to increase. In practice, textual format and reading practice adapt to minimise the effects of textual inflation. In the same way that electronic mail has made transactions more informal and immediate, so the range of message forms reflects varying levels of formality and consolidation of scientific knowledge. The more informal forms, such as bulletins and accelerated communications in *Perkin Transactions* and the *BIDS Chemical Science Index* are immediate but provisional in terms of scientific knowledge. More verifiable science is presented in high prestige refereed journals, such as *Trends in Pharmacology*, while 'popularised' articles appearing in *Nature* and *The New Scientist* represent a considerable time-lag between discovery and established knowledge. Within a list of research journals, therefore, a certain dynamic hierarchy can be seen to form which has a great deal to do with prestige and established scientific doctrine, as pointed out by Myers in his study of article rejection in biology (1990). In addition, part of the reason behind textual diversification lies in the fact that the majority of 'publication' is not public at all. Many texts are circulated to an exclusive number of specialists within an institution or between institutions and funding organisations, and this forms the 'grey literature' (Auger 1989). As information seeps out from institutions, the dividing line between exploratory bulletins, grant proposal reports

and polished refereed journals becomes blurred.

However, among the diverse and increasingly technological forms of communication used in science, the research article and abstract in a refereed journal are still considered to be the key elements not only in the raw give-and-take of specialised facts, but also in the maintenance of a community hierarchy and the dissemination of accepted ideology (Knorr-Cetina 1983:106, Swales 1990: 9-10, Kaplan and Grabe 1992: 214). In particular, the sociologists Latour and Woolgar (1972) have characterised scientific activity as predominantly the manufacture of written text, and postulated that written material is as much valued by the scientific community as the actual physical compounds they are manipulating in the laboratory. It is this material that provides the basis for exchange and individual promotion in the scientific group.

The US *Science Citation Index* (SCI: 1988) publishes a league table of 8 000 journals according to the highest impact factor, that is the relative number of times articles from the journal are cited elsewhere. Thus *Science, Cell* and *Nature* are in the top twenty, while *Mutation Research, Cell Differentiation and Development,* and*Ultramicroscopy* all fall within the last 600. This has important ramifications for our view of scientific discourse, since authors are expected by their institutions to publish articles in journals with the highest possible impact factors. Yet, as Ivor Williams ex-director of the Royal Society of Chemists has pointed out (Williams 1996), such quantification of science is problematic in that tertiary (accelerated) journals tend to be more cited than primary research. So while linguists such as Swales (1990) have established these texts as the 'traffic officers' in the flow of international science, we must regard research articles in the context of on-going research.

The research article has received much attention from researchers in ESP for its linguistic properties and rhetorical purposes (Swales 1990, Nwogu 1987) or as an original text for student summarising (Ventola 1991, Drury 1991) or as a prototypical scientific text for register analysis (Biber and Finegan 1991). These areas of research, with particular reference to work that has been carried out on the functions of research articles, are discussed in Chapter 3. Although context (translating lab notes, discussing with colleagues, submitting drafts) is essential in the understanding of the research article, it is nevertheless the product that is the interface with the wider scientific community and it is essential that its forms and linguistic properties be systematically set out.

1.3 Péritexte and the packaging of research articles.

The general concept of summarisation as a social activity, in whatever form, appears to play an increasingly central role in the management of an information-rich culture. From TV news bulletins to popularised science articles in the press, large amounts of information are being either cut out or reformulated into self-contained texts. In their work on specialist languages, Sager et al (1980) emphasize the "need for smaller packages of information" that are not simplifications but 'reconceptualisations' (1980:xix) with the particular property of being able to be reassembled by the specialist reader. They point out that in the processing of scientific messages, both the abstract and title play an auxiliary role between the sender and receiver of scientific information. Since information is growing in amount, researchers such as Jaime-Sisó (1993) have pointed out that titles are now fulfilling the functions of abstracts. We postulate that in turn abstracts function more as 'accelerated articles' in the communications sections of some journals.

Evidence of growth in the production of information in science is tangible in the statistics for the number of abstracts produced and processed. For one year in 1969 Maizell et al. (1976:6) report that the Chemical Abstracts Service (CAS), responsible for *Chemical Abstracts,* produced over 250 000 abstracts. For 1991 the senior editorial advisor for CAS estimates that the number of abstracts the service supplies has risen to an annual total of over 400 000, with an additional processing of 100 000 patents (Metanomski 1991). The workforce for such a large organisation has also doubled and numbers over 2000. Metanomski also notes that the abstracts are produced from an increasingly wide range of languages (25% other than English), Japanese being the most notable language for patents. These figures show that even for chemistry, where the market for new products is often based on reformulation of older ones, research activity is increasing exponentially and is also dependent on the changing fortunes of various nations' share of the research market. It is clear that changes in the role and form of scientific discourse can be seen as directly related to these factors.

While the number of journals of various kinds has been shown to increase (Swales 1990 and the British Library Document Supply Centre (Russon 1992,1993) both advance figures of around 100 000 periodicals) the number of abstracting journals and citation indexes has also increased, together with a diversification of the forms of abstract available. The *CAS* index of 40 000 monthly abstracts has extended the same service to an electronic commercial on-line counterpart *CAS Files*. Also *Medline* is typical of the type of indexing

service that is commonly available for scientists in the pharmaceutical sciences and for other disciplines. The system searches for several key words or phrases (*cancer* or *cancer cachexia*), or medical subject headings (*histology*) author's names and journals and then offers options on further more refined searches (cachexia: *-blood,* complications, *-diagnosis*) and provides abstracts taken directly from the original author. A typical search can pick up hundreds of 'hits'. It is then up to the researcher to sift through the titles and abstracts and then to find the relevant papers given the full reference. Other systems go further than this- *ADONIS* provides researchers in the medical sciences fully formatted articles including reference sections - an important source of texts for the corpus linguist. As university libraries attempt to save space and still maintain access to relevant journals, electronic indexes on internet represent the format of academic publishing of the future.

The physical format of chemical research is also evolving. Until now titles and references have only been given in indexes, but some journals now prefer graphic abstracts which can summarise an entire production process (or *synthesis*) of a new pharmaceutical product. The increasing miniaturisation of information, involving a more 'indexical' use of text has not only meant that the roles of abstracts and articles have changed but also that the presentation and selection of important information have become vital skills in the scientific research industry. We propose that this has specific consequences for the lexico-grammar of medicinal chemistry, and these are discussed in Chapter 2.

1.4 Remarks

This introduction has attempted to set out a justification for the study of scientific research and has associated this with its most privileged forms of communication (the research article and abstract) in the field of cancer research. A detailed survey of cancer research at Aston university is presented in Part II. In this section we describe a specific corpus of texts for phraseological analysis.

Three general questions were asked at the beginning of the introduction, and now these can be expanded and addressed by specific fields of research.

1.5 Primary question: What role do research articles have in scientific activity?

The relation of language to the notion of a separate knowledge structure or scientific substance has been the province of the information sciences. In particular the field of terminology has been professionally engaged in the rationalisation of language for science and has therefore built up a conceptual description of language that needs to be taken into account here. To the extent that terminologists and textlinguists share the general assumption that abstracts are 'extracts' representing the essential conceptual structure of an original text, then theories involving macrostructures of scientific text should also be evaluated in terms of their use to a description of linguistic practice in science. Developments in these three 'rational' approaches are overviewed in Chapter 2.

1.6 What role does the research article have in the creation of new science?

Having established research articles as vital tools in the hierarchy of science, not only do the relations between subject matter and language, but also the rhetorical functions of these types of text in terms of their professional context need to be addressed. Linguistic research may provide us with a mechanism for analysing the construction of scientific claims. As mentioned above, claim-building has been explored systematically in the field of English for Specific Purposes (ESP). The particular textual and linguistic forms of research articles have also be explored in detail and are summarised in Chapter 3.

Further, Halliday has provided a framework of discourse and grammar that describes the relation between language and scientific knowledge. Several methods of discourse analysis have been developed to describe the textual development of ideas, and this must be explored in any discussion of the textual creation of ideas. We discuss specific mechanisms of text creation in Chapter 4.

1.7 How does language function in extremely specialised contexts?

Recent studies in the field of computational corpus linguistics have attempted automatic characterisation of very large groups of texts in order to generalise and improve upon the findings of small scale studies carried out in ESP and discourse analysis. Since a phraseological and collocational view of language has emerged from terminology as well as from discourse analysis, it would be fruitful to explore the possibilities of phraseological

analysis in the context of corpus linguistics which has advanced a phraseological view of grammar but is yet to establish a descriptive tradition in applied areas. This is explored in Chapter 5 of the literature review.

LITERATURE REVIEW

2.0 CHAPTER TWO: TOPICAL REPRESENTATION IN SCIENCE.

2.1 The terminological view of science.

> Language is an instrument. Its concepts are instruments. Now perhaps one thinks that it can make no great difference which concepts we employ. As after all, it is possible to do physics in feet and inches as well as in metres and centimetres; the difference is one of convenience. But even this is not true if, for instance, calculations in some system of measurement demand more time and trouble than it is possible for us to give them. (Wittgenstein 1953, # 569)

We suggested above Firth's idea that 'topic', or what the text is 'about', may be indistinguishable from the use or functions of the text. For example, discourse topic can be seen as the political or professional implications of the text. Terminologists on the other hand set out a paradigm that attempts to pin down exact areas of scientific knowledge: a text for a terminologist is about the subject matter in the text, as far as Picht and Draskau (1985) put it. Godman and Payne (1981) set out a terminological view of language as an intermediary for abstract propositions, to be encoded by the sender and decoded by the addressee. Key to this is the idea of the *concept*:

> The term concept describes those elements, or related elements (i.e. a proposition) in the realm of thought that are expressed as a statement in language and form in the mind of the reader a identical set of related elements. (1981:24)

If any process is admitted, then the text may be seen as adjusting the way the scientific knowledge is mapped out by demonstrating how concepts are related or by elaborating new characteristics that need to be inscribed on the map. Godman and Payne certainly do not state that the map is fixed or that each element is not dependent on the others. While the folk vision of scientific text is about the description of 'things' that involves an inventory or nomenclature, traditional terminology sees the text as establishing a kaleidoscope of concepts that have abstract parts and interactions that must also be named. Mapping out knowledge by concepts, in the field of terminology, as opposed to setting out the semantics of words in lexicography (as Thomas distinguishes them 1993:44) is essentially a rationalisation of a specific body of knowledge, a paradigm or knowledge structure. While terminologists thus idealise the decontextualised theoretically fixed concept or *signifié* as knowledge, they offer a theory of reference that is shared with many of the concerns of the scientists for whom they work, and from whose number they often originate (Sager

1990:14). It is this view of language as a value based system that is of interest to the functional linguist: where terminology is appreciated by scientists as a problem that requires constant review. While the functional linguist is concerned with language in a complex social setting, the fundamental problems for terminologists are, as Picht and Draskau state (1985:38) (as established by Wüster (1933) in the fields of electronics and planned languages): the relationship between the concrete object and the abstract concept and the resources that languages may be given in order to express this relationship.

The ways in which the concept is expressed in language appear to be *ad hoc*. Picht and Draskau (1985) refer to set theory and logical as well as analogical relations between the real word and the idea as well as between the ideas themselves, making the conceptual world a complex multidimensional space. Sager et al. (1980) describe a suite of notions that terminologists use to classify different entities, as they term them, including objects (materials, instruments and products), properties (qualities of the materials involved), quantitative parameters of the objects (such as mass and velocity), processes (naming the applications of the materials) and the methods of such processes (1980:40). Sager et al. refer to these as language for special purposes (LSP) topics, and the question of their linguistic and non-linguistic representation in scientific publication, as well as the establishing of standards for their use, is key to the field of terminology (Sager 1990:8).

But many researchers in the fields of terminology and lexicography have questioned this approach, not least because it does not account for change, and have recently suggested an orientation that sees the concept as a mediation between competing scientific knowledge structures within society. The work of terminologists and lexicographers such as Pavel, Thomas, Godley and others who take a more phraseological view of terminology is set out in later sections. In the next section, the traditional tenets of terminology are explored, before a more detailed analysis of non-linguistic and linguistic aspects of terminology are set out.

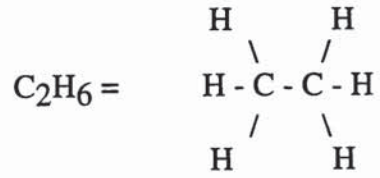
2.2 Non-linguistic forms of conceptual representation: 'Artificial language'.

While scientists cannot do without language for conventional reasons they point to the existence of artificial languages (Sager et al. 1980:40) as the fundamental objects and tools of scientific knowledge. Mathematical equations, formulæ, graphs, tables, figures, chemical structures, tree diagrams and other non-linguistic systems of representation

constitute semiotic systems in their own right for almost all scientific disciplines. One of the characteristics of artificial language is that it is "monofunctional" (1980:41), without discoursal function (signalling within the medium) or interactional function (signalling relations between the discourse participants), which would instead characterise natural language. Scientific language can thus be seen as the natural language form of artificial languages with the additional, and unique, functions of evaluation and interaction with the rest of the discourse community. But in terms of actual use, some of these resources constitute what we might call the 'indexical' use of language: figures are referred to by the text, and refer back into the text.

For the chemist, the graphic representation of chemical compounds as complicated as amino acids and chromosome rings is much more efficient in terms of recognition than their lengthy structural names. But even non-linguistic chemical symbol systems are redundant, and a cline emerges from the linguistic use of *trivial* or *commercial names* (where little or no structure is denoted, such as *ethanol*), *molecular names* (where ratios of elements are denoted: *carbondioxide)* and from *systematic* or *structural names* (where ratios and structural information are signalled as in *acyl-5-2'-deoxynucleoside*) to the more graphic *molecular formulæ* (ethane is C_2H_6), and *structural formulæ* as set out in figure (1) (Scott 1991:272).

Figure (1): Molecular and structural formulæ for ethane



Graphic representation not only denotes entities, but also processes and whole experiments. So useful are these reaction-graphs in structural chemistry, that, as mentioned earlier, researchers are often required to submit graphic abstracts which demonstrate at a glance the ingredients (substrates), steps involved (stages) and products of the chemical process (the synthetic story). In *Journal of the Chemical Society: Perkin Transactions* a sentence, often evaluative, is required with the graphic abstract, demonstrating the discourse functions of both systems. As can be seen in chemical writing of a highly specialised nature, the artificial language also merges with the natural language in the same text, until the mathematical symbols and formulæ are incorporated into the wording of the same paragraphs. Experimental sections in structural chemistry, for example, begin in natural

language and then embody concepts such as quantitative parameters, reaction times and percentage yields.

Apart from affecting the lexical level, non-verbal representation can be seen to interact with the cohesive and grammatical levels of the linguistic system. Throughout chemistry articles the names of chemicals and processes are replaced by numbers or figure references which retain their referents throughout the text, and so take on the function of text-long pronouns. The text becomes a cross-referenced instruction manual or index, which allows readers to process the text in a non-linear fashion (Roe 1977, Sinclair 1981:6, Tadros 1985:9, Tarasova 1993). Tarasova has demonstrated the interface of non-verbal and verbal forms of expression within the grammatical system. Non-verbal elements function as pre- and post-modifiers (*X ray, delta-V6* and so on) as well as nominal groups (*as T goes to* T_S) (1993:47,56). She also finds that the occurrence of prepositions with non-verbal elements is particularly characteristic of the scientific genre for example, *at phenon energies of 5.6 eV*.

It is not simply a question of 'translating' from graphic to linguistic or back again. Each choice in the system, according to Firth's approach, has some communicative motivation. Given the wide facility for reformulation, to what extent can linguists predict the factors that lead to the preference of one of many types of method of representation over another? What makes the complex format of chemistry texts interesting is the fact that specialist writing is seen as the interface between novel ideas and an already established scientific paradigm. Sager et al. claim that non-language can be simply divided from language because of natural language's discursive properties:

> The borderline between natural language and artificial language can be said to lie at the point where natural language loses the ability to be its own metalanguage. (Sager et al. 1980:42)

Yet Godley (1993) has demonstrated that representation in chemistry is not as objective or as universal as it superficially appears. The representation of benzene rings, for example, as the basic structure of organic compounds is not accepted by all journals and abstracting services: when writing for *CAS* the researcher is required not only to reformulate diagrams to fit in with the editorial view of the matter, but also to change structural terms (which are supposed to reflect three-dimensional structures by numbering according to positions on the benzene ring). Other disputes exist about whether metals should, in linguistic terms, be the 'heads' of noun groups (thus meaning that valency is reflected in modifiers) or the other

way round. Ideology, paradigmatic correctness and convention means that there is a degree of fluidity even within this seemingly hermetic system.

The amount of redundancy in chemical nomenclature and graphic representation is therefore a key indicator of the complex discoursal processes that are involved in the writing up of science. This leads us back to the first of the preliminary questions raised in Chapter 1. The question of representation is fundamental to the fields of terminology and lexicography and is an area of large polemic even within the fields of medicinal chemistry and cancer research themselves, as Godley (1993) has suggested. Before looking at the linguistic side of terminology, we can already posit that scientific writing is a rhetorical continuum stretching from conventions of textual format through language into the cohesive system and the non-verbal system as well. The whole works as a Firthian system, with different levels of acceptable rhetorical strength being applied to different levels according to the genre (a linguistic abstract in a maths paper) and the discipline (the entirely linguistic representation of the mind in psychology).

2.3 Linguistic forms of conceptual representation: The 'special' language.

The term 'languages for special purposes' has few differences in essence with 'specific' preferred by Anglo-american workers in ESP. But the choice of the term 'special' does indicate a closer relationship to the topic of the language specialism and a generally terminological flavour to the research orientation of terminologists and information scientists. For the terminologist, the difference between 'language for general purposes' (LGP) and 'language for specific purposes' (LSP) is expressed by Picht and Draskau in terms of 'abstraction', the degree of abstract linguistic specialisation in which any topic may be discussed:

> The level of abstraction is defined as the ratio of abstracts to concretes. 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)

Picht and Draskau see the difference between LSP and LGP as a cline of expression of essential content. But 'essential' here does not mean 'the same'. Abstraction involves the introduction of an increased level of generality, so that while *Cologne cathedral* indicates a specific real world object (denoted by a *name*) the concept *cathedral* is abstracted away from reality to a generic idea (denoted by a *term*). Picht and Draskau note that abstraction is reflected in the characteristic nominal style of the LSP, while the LGP has "a zero level of

specialisation". In the same way as artificial languages described by Sager et al., Picht and Draskau characterise LSPs as "monofunctional". LSPs are thus negatively defined in that they cannot be 'picked up' by the lay person, are restricted to exclusive groups and are seen as non-essential in the wider community (1985:10-11).

A functionalist, Firthian account of language would not see 'special language' in terms of degrees of specialisation- after all all human activity is 'special', the only difference being, Halliday and Martin (1993) argue, that science has a superior cultural position. Similarly, the monofunctional hypothesis about 'special languages' does not correspond to Godley's (1993) observations of language-like redundancy in the terminological system of chemistry. A more flexible view comes from Pavel's work on terminological vocabulary. She puts special emphasis on the idea of a world-vision being the largely differentiating factor, rather than abstraction or function:

> La langue de spécialité est un sous-ensemble de la langue générale [LG] qui sert à la transmission du savoir relevant d'un champ d'expérience particulier. Elle n'existe qu'en partageant la grammaire LG et une partie de son inventaire lexico-sémantique (morphèmes, mots, syntagmes et règles combinatoires) mais en fait un usage sélectif et créatif... Les diverses langues de spécialité appartenant à une même langue générale reflètent la vision du monde propre à la civilisation dont la LG est issue. (Pavel 1993:1)

To this definition, Pavel adds that each discipline has its own vision of the phenomena that it is studying. Before discussing Pavel's position in more depth, it would be useful to study terminological views of text that still claim the existence of 'special' languages, but attempt a functional approach.

Sager, Dungworth and McDonald set out a 'specialist' view of texts within science. Science is considered to be more dynamic than other specialised human activities such as crafts and technology, and as it innovates it makes a high demand on the terminological resources of language (1980:xviii). Basing their discussion on the categories proposed by functional linguists such as Hjelmslev, Bühler and Halliday, Sager et al. claim that there are three types of discourse that correspond approximately to Halliday's textual, interpersonal and ideational metafunctions (1980:85). *Metalinguistic discourse* with extra and intertextual comment is, claim Sager et al., untypical of scientific texts, and as mentioned above, is a resource that appears to fade away as the language becomes increasingly graphic and conceptual. *Perceptual discourse* concerns reference to the immediate physical and temporal context of the text itself. Finally, *conceptual discourse* is concerned with environmental reference beyond the physical reference of the text into the abstract conceptual world of

scientific knowledge. Such a relationship between language and knowledge is expressed in terms of *conceptualisation* (1980:xx).:

> ...we assume knowledge structure to consist of a multidimensional hyperspace with orthogonal axes in which any concept can be identified uniquely by reference to its co-ordinates along each axis. (1980:70)

As a consequence of this, the terminologist's task is to establish formally defined standards of use of linguistic resources in accordance with special knowledge, and its attendant artificial languages (Sager et al. 1980:40). In this way the language of medicinal chemistry constitutes a typical special language, and as such can be contrasted to the general language where informal norms of usage hold sway, and where terminological problems, such as neologisms, are in constant flux. In addition, both special language and special knowledge are contrasted to registers, such as the language of journalism or administration which are instead destined for interaction between different discourse communities (1980:4).

The effects of terminology in the wider linguistic system can be seen as linguistic resources are manipulated in an attempt to represent semantic universals and semantic relations within a highly abstract conceptual space. Thus words are "pressed into service" as conventionally agreed terms by means of adapted resources such as *conversion* (e.g. clone from noun to verb) or *derivation* (Sager et al. 1980:15,78). The linguistic mechanism of derivation is seen in the special adaptations of natural language affixes using subject specific dependent morphemes in chemistry, such as *-ous* (indicating less oxygen bonds as in *sulphurous acid* H_2SO_3) *-ic* (indicating more oxygen as in *sulphuric acid* H_2SO_4) or *-ate* (as in *sulphate* indicating a compound that includes either the SO_4 ion or SO_3 ion.) (Scott 1991:272-278). Derivation may also typically take the form of what we might term 'agglutination' where independent elements are compounded, or compounding by juxtaposition but leaving a space or hyphen between individual elements, essentially the formation of new complex nominals. Sager et al. (1980:270-272) classify compound nouns into four syntactic categories - adjectival compounds (*compressive force*), operation compounds (from *a change of temperature* to *temperature change* where subject/object/instrument is often obscured), deverbal compounds (*dust collection*) and verbal nouns (*air-conditioning, town planning*). On the semantic level, Sager et al. (1980:268-269) classify nominal two word compounds into the following categories:

a) the head is compared to the modifier: *ethane-type interaction.*
b) the head is made of a specified material: *oil film.*
c) the head has a new property: *low octane.*
d) the head has a specific use: *cutting tool.*
e) the head is associated with its product or origin: *malt beer.*
f) the head operates on the modifier: *enzyme reactivator.*
g) the head operates as specified by the modifier: *sliding key.*
h) the head is part of the modifier: *pedestal cap.*
i) the head is identified by the modifier: *gold standard.*
j) the head takes place at the modifier: *cytokine tumour.*

Affixation and compounding therefore provide a powerful but conceptually *ad hoc* terminological resource as can be seen in the 750 000 compounds and 4 000 000 terms (including affixes and suffixes) in organic chemistry and 30 000 in inorganic chemistry (Sager et al. 1980:230).

It is worth noting here, as mentioned in the discussion on non-linguistic representation, that a particular complexity of chemistry that is brought into the system of compounds is the idea of spatial representation, as Godley (1993) has pointed out. Thus a chemist may be able to decide on which nodes of a ring a functional compound is situated by using figures in between each functional group. Hence the structural name *5'-acetyl-3'-thiophenylthymidine* includes figures that are suffixed by an apostrophe in order to distinguish them from molecular, tabular or other notations.

Key to the fields of terminology and lexicography is the *definition*, 'the verbal description of a concept' (Picht and Draskau 1985:65). Systems of definitions present a complex area of analysis 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 of *extension* (an entity's associated purpose or functions) (1985:47). In addition, a highly defined concept in one scientific discipline will be interpreted as pragmatically different in another. For example, the molecule $FeCl_3$ is important for electricians as well as textile technologists, but has a different meaning (of *extension*) in both fields (Sager et al. 1980:72). For cancer research, we note later that researchers even have a different perspective about concepts as central as cancer.

From the complicated semantic picture of conceptualisation, the essential idea of *reconceptualisation* emerges which involves the changing functional perspective of concepts and terms. Sager et al. do not set out an entirely static view of terminology: the basis of

change in their view is based on manipulating the representation of concepts. But against their vision of the clearly defined division between the metalinguistic world of the 'general language' and the conceptual world of the 'special language' we turn to a view of language as not only the recipient of knowledge but as a primary tool in the changes of paradigm as envisioned by Kuhn (1962).

2.4 Terminological change and Pavel's *LSP collocations.*

Béjoint (1988:365) sets out to challenge the fixedness of terminology and conceptualisation. He makes the observation that as one's viewpoint changes, so the conceptual constellations change, as in the meaning of 'star' in the following two sentences:

> 1- You can't see the stars because of the sun.
> 2-The sun is a star.

He also sets out (1988:357-359) a set of characteristics of scientific and technical words that are often claimed to hold true by terminologists:

-Scientific terms follow a chain of definition down to LGP items.
-Scientific terms enjoy an "absence of ambiguity in context and out of context." (1988:358)
-Scientific terms have no figurative or metaphorical meanings.
-Scientific terms have origins that can be definitely traced.

Yet 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 can be seen as totally un-metaphorical. His key point, however, is that the process of terminological definition is circular, and this touches at the heart of the rational image of science. These comments are echoed by Godman and Payne (1981:24), who point out that the idealised knowledge structure is exposed to the flux and uncertainty that is prevalent in the general language. Thus the general and special languages cannot be separated; instead they blend into each other since the change of one concept in the knowledge structure would affect the position of the others in the knowledge structure of the LGP or the LSP. As Godman and Payne put it:

> Each term is dependent for a full appreciation of its meaning on the meaning of the other terms of the group. (1981:28)

This challenges the underlying assumption that greater precision, for the terminologist, can be defined out of context, as the terminological commissions of the *International Standards Organisation* (ISO) might suggest.

To modify the model of terminology, the Canadian terminologist Pavel has postulated a dynamic view that change comes from within the textual process of scientific writing. This view of terminology stems from her work on the creation of a collocational terminology of 'systemics': a post-fractal philosophy that brings together traditionally disparate disciplines, such as linguistics, physics and economics, where a number of researchers recognise the complex dynamic systems in operation (Pavel and Boileau 1994:1). Since fractal imagery is largely adapted as metaphor from everyday language, the terminology is particularly transparent to the non-initiate. Pavel sees the effects of this on the shape of the conceptual network of fractal science. The effect of metaphor on whole areas of thought has always been perceived as a motor of innovation. For example philosophy has been revolutionised on the basis of simple metaphors such as Aristotle's Substitutions and Comparisons or Austin's Speech Acts, as Koch (1991:290) has noted.

Pavel considers Picht's (1990) ideas of conceptualisation in the information sciences as a constantly changing cycle of novel formulations expressing metaphors that are transferred gradually to the main knowledge structure of the scientific community. Thus the conceptual knowledge structure, even when agreed upon, is set to change:

> ...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 argues that new formulations, which she terms *phraseology*, effectively reconstruct the terminological knowledge structure. As new phrases become neologisms and accepted terms, these in turn bring their own suite of associated terms, sometimes from different disciplines. Pavel refers to these as *LSP collocations* (1993a:29). She recalls the terminologisation of otherwise 'general language' terms, such as the semantic field of the theatre in one model of artificial intelligence (namely: Schank and Abelson 1977), including terms such as 'scripts', 'actors', 'thematic roles', 'frames' and 'props' where the conceptualisation of the brain is of "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 that make up their use in their original context. Pavel describes these processes in terms being initiated, negotiated and finally accepted by the wider scientific community (:27):

> ...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)

Reversing the process, as scientific discovery is disseminated into popular culture, such reformulation brings with it changes in the accompanying belief systems, or conceptual systems in the language of the terminologist. This to and fro of concepts, with attendant belief structures, is encapsulated by what Pavel terms the *thematic proposition* (1993a:30). Pavel describes the formal aspects of LSP collocations as a linguistic resource in her work on phraseology (set out below), but she does not distinguish them from thematic propositions. One may assume that the LSP collocations are the phraseological structures that accompany a concept, while the thematic proposition denotes the entire process as the neologism becomes embedded in the scientific knowledge structure. This embedding of thematic propositions in turn accounts for the progressive change in the rest of the knowledge structure that Kuhn has established as 'paradigm shifts', where concepts cannot maintain the same meaning when other parts of the system change. Thus Pavel's view of terminology as a system is synonymous with Firth's idea of systemic meaning:

> Thematic propositions incessantly question and undermine the concepts designed to grasp them. [...] Like all dissipative systems, thematic propositions acquire new features, forfeit previous ones, restructure internally and diffuse outwardly. (1993a:30)

Pavel and Boileau's (1994) dictionary of fractal terminology contains typical collocations and synonyms of main entries. A noun followed by a verb group (*this instrument cuts paper*) can be reformulated by embedding in a noun group (*the instrument that cuts paper*) (1993b:4). Pavel distinguishes between phraseological units (e.g. *represent in the form of a fractal*) and terminological units (e.g. *fractalise*). Terminological units may be classed according to phraseological combinability (fixed, restricted, free), internal variability (by synonyms of key elements), degree of compaction (how much the key element can be isolated or reformulated), frequency and specialisation (1993b:6). The difference between this and phraseological unit is one of degree, although Pavel suggests that the terminological unit is more readily accessible in the form of nominal, adjectival and verbal groups, and certain semantic networks emerge that are specific to the model in question. For example in fractal imagery N+N groups display inclusion (*particle-cluster*), N+Adj groups display gradual superordinates (*chiral chemical compound*), N+V collocations display specialisation (*the product crystallises* (transitive sense only)) and V+N predicates display directionality (*conserve scale*) (1993b: 5). For the scientist therefore, such a sorting

of collocations according to terminological information is more useful than grammatical information in any collocational description of terminology, as Béjoint and Thoiron point out:

> 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)

Pavel's LSP collocations provide this thesis with a powerful metaphor and a useful link between the rational scientific approach of terminology and neo-Firthian approaches to language which have always held the collocational principle of reformulation as a cultural as well as linguistic resource. One issue remains to be explored from the cognitive point of view, that of the idea of an independent subject matter. Having established that phraseology, or reformulation, plays a role in relating 'real world' meaning to linguistic form, we must go one step beyond the lexical level to look at the concept of subject matter as a function of text.

CHAPTER THREE: LINGUISTIC MODELS OF SCIENTIFIC WRITING

3.0 Purpose and rhetoric in scientific texts.

> 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 Firthian theory of language functions has already been linked with the field of English for Specific Purposes in the introduction to this thesis. Having discussed the terminological approach to language in terms of concepts, it is now essential to discuss the goals, needs and practices of scientists in their working environment, a concept referred to as *discourse* in preference to a decontextualised concept of 'language' or a propositional macrostructure of 'text' (as proposed by the textlinguists, Van Dijk and Kintsch 1977).

3.1 Variety in discourse.

A functional model of language (Firth 1957, Halliday 1966, Berry 1977) postulates three metafunctions of language communication: textual, ideational (knowledge-related) and interpersonal (pragmatic) components. In terms of scientific writing, we have already seen that the textual element interacts with the knowledge structure of a discipline in the form of terminological devices: a scientific variety of language with its tendency to conceptualise and form complex nominal groups, with at the very abstract level special and artificial languages. The textual component also interacts with interactional criteria in terms of 'genres' (Swales 1990) 'text types' (de Beaugrande and Dressler 1981:85) or 'special text units' (Sager et al. 1981), characterised by their essential purpose of communication. At the same time the interpersonal metafunction encodes the knowledge structure according to group usage in terms of specific lexis and terminology.

Aside from the lexical and grammatical instantiations of these functions (discussed in terms of discourse analysis, later) the metafunctions may be seen as determining factors in the final form of the text. Thus Sager et al (1980) break down the primary functions into variable categories such as participants' status, which determines the textual frame in which the knowledge structure is to be fitted. 'Aspect' is one element of participant status: the use

to which the text is to be put (administrative, pedagogical, descriptive) (Sager et al 1980:102). 'Mode' is the other element at the level of status: the level of formality and advanced planning involved. The participants' knowledge is the second half of the topic equation, involving 'level' of reference (from specialised to popular) and 'field' (from the very broad field of physics to the narrower field of nuclear physics). Sager et al. (1980:120) claim that these four participant characteristics manifest themselves in five propotypical categories of 'special text':

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.

Purpose appears to be a major defining factor which accounts for the difference in form of these genres. Sager et al. (1980:125) use four categories of Searle's (1969) speech acts to describe intentionality in special texts: informative, evaluative, directive and phatic. However, Nystrand (1986) has argued that purpose cannot account for linguistic form alone: it has to be accompanied by a system of *reciprocity*. Nystrand states:

> Writing, though clearly monologic as an activity, is nonetheless dialogic in its communicative structure. Each point at which the skilled writer chooses one example rather than another, one term rather than another, certain comparisons rather than others etc, is ultimately arbitrated not only by what the writer has to say but also by the need of his or her reader to understand. (1986:36)

Thus exactly shared or mutual knowledge is not necessary for understanding, as long as writer and reader can find a common *frame of reference* (1986:54). Thus comprehension, for Nystrand, is the sum of frames of reference:

> Comprehension is always affected by previous text and expected text, as well as the nonverbal context in which the text is situated. Comprehension is the process whereby words emerge as meaningful constructs from otherwise empty perceptual forms. (Nystrand 1986:58)

Thus the meaning of a part of a text is couched in the rest of the text. Local meaning cannot be independent of a researcher's previous reading and experience. Nystrand claims that the reason why readers can skip and skim through text is that can predict what the text says. Sinclair (1993) has argued a similar position ('posture' in Chapter 4 below), arguing that

local signals about a sentence's relation with past and future text are a fundamental resource of discourse. If prediction is an important factor in narrowing down coherence relations in text, then the discourse community of scientists should find textual conventions an important resource in research writing, as discussed below.

3.2 Conventions in genre analysis.

Swales has established a methodology of 'quick and dirty' analysis to determine conventions in writing. This involves foregoing the theoretical niceties of existing models by addressing the needs of specific professional groups who, in the case of scientists, consist of researchers in their roles as writers, editors and readers. These groups are termed Discourse Communities "...sociorhetorical networks that form in order to work towards sets of common goals." (1990:9). The group may consist of individuals with different research interests and specialisms, but their long term goals will be agreed and supported by mechanisms to enable information to be shared by members of the group. These mechanisms include "control of technical vocabulary" and the establishment of a hierarchy of expertise (Swales 1990:32).

The common goals that constitute the internal cohesion of the discourse community are realised through language events and resources. Swales claims that his analysis of textual genres ultimately stems from Propp's (1923) 'Morphology of the Folktale'. Folktales work because their readers are familiar with the conventional formal schemata, so readers expect such events as the damsel being in distress or the protagonists living happily ever after. Such Genres are defined as "...the properties of discourse communities... classes of communicative events which typically possess the features of stability, move recognition and so on." (1990:9). Swales thus sees the genre as means to an end, fulfilling a definite set of communicative purposes (entertaining the audience, selling scientific ideas) and very often conventionally labelled by the group (fairy tales, review articles). But an important point is that genre is defined by its use alongside other genres, not just by its internal linguistic features.

In genre analysis, unlike Biber's (1988) register analysis, the principle is that grammatical features which are superficially the same as general language features function in accordance with the linguistic practices of the discourse community. Swales thus contrasts genre with register (1990:41) which is the general grouping of linguistic features that characterise a certain text, and in the context of this thesis could be taken to mean 'the

language of cancer research'. Swales (1981c) demonstrates this in his analysis of the past participle in technical English. Swales finds that its major function is to bring the reader's attention to non-linguistic text. This use is different to traditional descriptions of the general language, with the postmodifer referring to given information or with some proximity to a non-verbal part of the text, as in *the curve shown, the list given*. Swales refers to this as 'discourse coherence' (1981c:45). Similarly premodifiers such as *a given reaction* function in a similar way to pseudo-determiners as in *a certain reaction.*

Swales' main contribution to genre analysis has been his characterisation of the rhetorical structure of article introductions, with the view that rhetorical sections (introductions, methods and so on) represent the fundamental conventional framework of scientific articles. His research is presented in the next section, in a general synopsis of research on all the relevant rhetorical sections of the research article genre.

3.3 The research article as a working genre.

In the context of the massive flow of written data in science, Swales identifies 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 by the discourse community. In the case of the research article this includes graphical and textual format as well as devices for academic accreditation and citation (Swales 1990:6). At the discourse level, Swales identifies a conventional stereotypical rhetorical structure that is analogous if not formally similar 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 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 and sub-moves or steps (1990:137). Since moves are rhetorical in nature they represent a summarisation *in essence* of recurring rhetorical structures that the argument of a text can go through.

The first move, '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 that are pertinent to it:

M1 Establish a territory:

S1 Claim centrality.
S2 Make a topic generalisation.
S3 Review items of previous research.

Since the three steps leave more to interpretation by involving decreasingly explicit explanation of the discourse topic, Swales characterises the direction of the global information structure in move 1 as "declining rhetorical effort" (1990:141). 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 about the 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, a reporting phrase (*was found to be*) or reporting verbs (*show, demonstrate, suggest*), and a 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, by reducing the significance of or enhancing the old model:

M2 Establish a niche.

S1 Counter claim.
S2 Indicate a gap.
S3 Raise a question.
S4 Continuing tradition.

Since the relationship between claim and model is made more implicit and less sharp by the later steps of move 2 (in step 4 there is no counter-claim), Swales identifies the rhetorical direction as decreasingly strong (1990:141). The linguistic characteristics of move 2 involve references to the negative effects of previous methods in conjunction with grammatical negatives or conjunctions of adversity (*However, few*) and lexical negatives (*fails to, is inconclusive*) (1990:155). The weaker 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 discourse topic on to occupying the research gap established in the first two moves:

M3 Occupy the Niche.

S1 Outlining purposes.
S2 Announcing present research.
S3 Announcing principal findings.
S4 Indicating Research Article structure.

According to Swales (1990:160) the linguistic features of this move involve a lack of references to previous research, explicit deictic references to the text: (*the present authors, in this paper*) and a prevalent present tense (1990:160). By stating the aims of the new research and exploring the detail of research, move 3 takes the rhetorical direction into the 'present' text with increasing explicitness (1990:141).

With this influential model in mind, we summarise below linguistic work on research articles as a whole, and then research on specific rhetorical sections of research articles.

3.4 The research article as a whole.

Swales (1990:134) states that the research article is divided into different functional sections that make use of different linguistic resources. Atkinson (1992) has traced the development of the traditional scientific paper and the development of rhetorical sections which we shall codify as TAIMRD(R): (Title - Abstract - Introduction - Methods -Results - Discussion - References) with the IMRD sections being seen as the core of the paper and receiving the most attention from researchers. Research on abstracts has been of a different nature, and research on titles, methods and results sections has hardly been undertaken.

Bruce (1983) has found that the IMRD structure in medical articles corresponds to the natural process of inductive enquiry. However, Blanton (1982) has suggested that real rhetorical divisions may not be formally the same as the visible sections of an article, distinguishing between discourse hierarchy as the conventional "observable organisation" of scientific reports, and rhetorical strategy as the "underlying intuitive development" responsible for the internal logic of each rhetorical section and as scientific reports. This distinction may help us to describe the use of rhetorical sections (such as combined Results-discussion sections or abstracts) which sometimes function as coherent, self-contained texts. In the analysis of posture in Chapter 9 we find evidence to support this.

Some studies have analysed the language of different sections, others have analysed linguistic features related to the discourse of the article as a whole. Studies of the research

article as a whole have established that lexicogrammatical features (often verbal tense and mood) are connected to specific rhetorical functions, such that statements about the use of the passive or authorial comment depend entirely on the subject matter, domain and rhetorical section of the research article.

Oster (1981) finds that verb tense is involved with signalling that information is about to be presented later on (1981:87). He also finds that non-finite verbs tend to be used for attribution as pre-modifiers (*tumor-derived factors in...*) or in non-finite clauses (*in supplying fatty acids, lipid mobilization...*). Sager et al. (1980:218) find that when non-finites are in rhematic end-of-sentence position, they signal a result as in: *...leaving all the gears exposed.* Wingard (1981) analyses verb usage in 15 medical texts, showing that up to 40% of verb uses are passive as opposed to 60% active, and while the simple present is the most frequent form (28-40%), 64-78% of verb uses are non-finite (70-80% of which in past participle post-modifying noun phrases). In 20 MSc theses, Hanania and Akhtar (1985) obtain different results, showing a preponderant use of the past tense in methods sections (usually in conjunction with the passive). Malcolm (1987) makes a distinction between rhetorical constraints on grammar and rhetorical choice. Because they are obligatory, an authors' use of the present for generalisations, the past for specific experiments and the present perfect for footnotes are all constraints. On the other hand, the use of the present or the past in describing previous research as either specific or theoretical, and the use of the present or the present perfect to accept as given or to distance oneself are all strategic rhetorical choices available to the author (1987:38-40). Gunawardena (1989) discusses the multifunctionality of tenses such as the 'retrospective' present and the 'inclusive' present. Tense, therefore, cannot be seen in terms of time but also in terms of authorial evaluation of the information he/she is setting out. Thompson and Yiyun (1991) further classified reporting verbs in research articles, distinguishing between author's stance (where evaluation goes from praising to negative) and writer's stance (where statements are accepted as fact or non-fact).

Research on 'phraseology' or aspects of writing not concerned with tense are less frequent and very specific in scope. Master (1991) has found that inanimate subject nouns (*shuttle, particle*) are more likely to have active verbs than passive verbs, which are more likely to be verbs of causal processes (*cause, affect, prevent*) than reporting verbs (*show, indicate, suggest*)(a distinction echoed in the PSC corpus for many patterns, as described later). Dubois (1981) explores noun-phrase embedding in research articles. Abraham (1992) finds that the use of *because of* which signals given information accounts for 41% of the

occurrences of *because* (a signal of new information) in scientific writing as opposed to 6% in spoken discourse. 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 comparatives and phrases with *show*) and implications (involving a high degree of modality). Thyman (1981) proposes that the description of nonlinear (simultaneous) events in scientific writing has led to the evolution of specific functions of cohesive devices, such as the classifying and defining function of *this*. We later see that this is a key item in the process of reformulation.

3.41 Titles in research articles.

The title can be considered a 'rhetorical section' by definition, although little research on IMRD sections has found the title useful for comparison with the rest of the article. Since most studies involve corpora of less than 50 articles, the number of titles would not yield much data.

However, Jaime-Sisó (1993) has carried out a diachronic corpus analysis of over 2 000 journal titles from 6 fields of medicine from the electronic indexing service, MEDLINE. Since 1980, Jaime-Sisó finds that there has been an increasing frequency of titles (from 0% to 40% of a yearly sample from the same journals) that involve clauses *(Dietary fish oil delays puberty in female rats*), a structure she terms 'new titles'. By comparison with the types of journals and papers that the new titles occur in, new titles can be said to characterise areas of science where there is an increasing number of active researchers (namely the new field of developmental biology) and where the journal is consistently high on the impact factor scale. Jaime-Sisó finds that the types of verbs involved in the clauses (*contribute to, is required for, contains*) oblige the author to justify the novel results elsewhere in the article, the role of the title effectively becoming a promissory notice of results.

The point here is that linguistic change reflects the changing role of the title in terms of its environment, which might involve the growing use of graphic abstracts, increasing independence of the title and abstract as 'stand-alone' text types, and the non-linear use of texts.

3.42 Abstracts in research articles.

The abstract's allure for the linguist lies in its function as a text presenting the 'essence' of a non-linguistic scientific event, responsible for representing and replacing the 'original article', and at the same time selling itself to a demanding public of editors and readers. As a result, more research on abstracts has been undertaken than on other sections - but largely in the information sciences and in fields such as textlinguistics.

Most research centres around two basic types of abstract. The informative abstract introduces the article's main ideas and explains the essential points of the original article. The indicative abstract reformulates each major rhetorical section of the article following the progression of the article as closely as possible. Of the two, the involvement of the writer should in theory be more active in the informative abstract (Cleveland and Cleveland 1983:4).

Most research, especially in the fields of the information sciences and documentation studies, has examined indexing-abstracts where the abstract is produced by a professional abstractor or abstracting service. Baker et al. (1980) have analysed the role of the professional abstractors at *CAS*. Bernier (1985) and Craven (1965) have analysed the syntax of 'terse literature' including abstracts. Weil et al. (1963), Cleveland and Cleveland (1983), Cremmins (1982) and Memet (1986) have set out practical guidelines for the professional abstractor. Dronberger and Kronitz (1975) and Reder and Anderson (1980) have studied abstract readability as a function of vocabulary of indexing-abstracts and Fidel (1986) has analysed vocabulary differences between indexing-abstracts and conclusion sections. Meyer (1988) and Gibson (1992) have set out the functional linguistic features of trainee abstractors' successful and unsuccessful abstracts. Rush et al. (1971), Pollock and Zamora (1975) and Sharp (1989) have analysed the possibility of using professional reading procedures to automatically produce indexing abstracts. Khurshid (1979), Polskaya (1986) and Raya (1986) have all examined index abstracts from theoretical information science viewpoints.

Related to professional abstracting, there has been much linguistic research on summarisation following Van Dijk and Kintsch's (1983) propositional textgrammar and De Beaugrande and Dresslers' (1981) studies on topic-summaries formed by the matching of textual patterns. However, most of this research has been applied to analyses of student summaries (Frank 1971, Fløttum 1985, Sherrard 1989) although Gopnik (1972) has set out a textgrammar of science abstracts. She categorises minimal processes similar to the

textgrammarian's propositional 'macrorules'. Typical of these studies, Johns and Meyes (1990) find that lacking the background knowledge of a specialist field, non-expert summarisers delete the wrong information and construct propositions on false premises.

Author abstracts appearing in refereed journals have been analysed on a smaller scale by more rhetorically oriented studies; including Borko and Chatman's (1963) guidelines for authors. Endres-Niggemeyer (1985) emphasises the lack of influence of abstracting guidelines on the rhetorical framework of author-abstracts. Buxton and Meadows (1978) set out common points of information contained in chemistry abstracts. Harris (1985) has studied authorial comment and stance in scientific abstracts and Sastri (1968) has analysed prepositions in chemical abstracts. King (1976) has set out the typical vocabulary profile of author abstracts in a study related to Fidel's (1986). We note here that the corpus analysis we conduct in Chapter 11 challenges some of these findings.

Unfortunately, few articles that deal with differences between IMRD sections consider abstracts at the same time, and vice versa. Gläser (1991) has argued that the abstract is a separate genre rather than a rhetorical section, and points to its condensed presentation of content (with high compaction of nominal groups) and lack of deictic reference or stylistic devices. Grätz (1985) has found that most abstracts follow the rhetorical structure of IMRD sections. However, Endres-Niggemeyer's (1985) comparison of IMRD instructions for authors and actual rhetorical structures of abstracts and articles in chemistry journals suggests that writers' sections correspond to the perceived particular needs of the reader rather than the journals' style guides. She proposes instead conceptual text types situated around topical poles, such as the *overview* and *model building* abstracts or *practice oriented* and *theory-descriptive* abstracts (1985:45). Similarly, Salager-Meyer (1990b) finds that 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 logical signalling and cohesive devices are usually absent in abstracts (1990b:378).

Salager-Meyer (1992) also analyses verb tense and voice usage and modality in the 84 abstracts of 49 research papers, 21 reviews and 14 case reports, all subdivided into clinical, basic, epidemiologic and operative types. The active past tense is the most frequent tense (51% across all types) and corresponds with purpose, results, methods and case presentation moves. The past passive is particularly prevalent in the methods move, indicating an obligation to use it. In the purpose and conclusion moves, on the other hand

Salager-Mayer states 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 is also prevalent in the data synthesis move of review article abstracts, which Salager-Meyer equates with the conclusion move in other abstracts. The present perfect also has the dual 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 statement of the problem and data synthesis moves, 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, *may*, indicating high probability of claims usually in the conclusion move, *can* being associated with data synthesis, and *should* heavily outweighing other modals in the recommendation move (1992:105).

On the level of lexis, 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 first paragraphs of the 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. Gibson (1991) and Drury (1991) have both demonstrated that non-author abstracts that are perceived to be successful tend to have topical themes as opposed to textual and interpersonal themes. Drury (1991) finds that rather than simplifying texts abstracts need to render them more abstract and technical (1991:436). The successful summariser reduces the amount of relational and embedded material processes from the original, introducing more material processes at the rank of clause (1991:447). This is mirrored by increasing lexical density and use of grammatical metaphor in the successful summaries (1991:448).

Nwogu (1989) has analysed cohesion, thematic progression and Swales' (1981a) system of moves in 15 medical research articles, 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 *purpose* and *methods* moves: presenting background information, reviewing related research, describing data-collection, describing experimental procedure, highlighting overall research outcomes, explaining specific research outcomes) (1989:171). The moves 'describing the data-analysis procedure' and 'indicating non-consistent outcomes' do not occur in the abstract (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 leads to complex clause structures and a greater sense of compaction (1989:180).

Kretzenbacher (1990) has conducted a corpus analysis of 20 abstracts and their original academic research articles in German (88 000 tokens). He finds that abstracts have a strong nominal profile, with a significantly higher noun-per-sentence ratio, more 'verbal substantives' in German (which are usually marked by the equivalent abstract noun suffixes *-ness*, *-ity* etc. in English), and more nominal compounds than the original article (1990:56-67). Articles are marked verbally: the majority with significantly more finite verbs, although descriptive abstracts have a majority of passive forms. Contrary to previous research, only 8 of the 20 articles have relatively more modals than their abstracts. Abstracts are found to have a slightly lower word per sentence ratio than texts, (23.8 to 24.62) which is still high in comparison with other German genres (1990:86). Kretzenbacher explains this by stating that articles use greater use of parataxis and hypotaxis while abstracts have relatively more use of embedded clauses. Also, abstracts tend to use nominal groups and finite verbs as attributive elements of clauses, a typical German construction (1990:101). Kretzenbacher also finds that abstracts have relatively more genitive attributes (with *von*) and definite articles while articles have more infinitives, anaphoric reference, and personal deictic reference.

3.43 Introduction sections in research articles.

Apart from Swales' (1990) analysis of introductions set out above, West (1986) 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.

Similarity between the rhetorical functions and grammatical features of introduction sections and discussion sections has been often noted. Gunawardena's (1989) analysis of 10 biology and biochemistry articles shows that the present perfect is particularly prevalent in introduction and discussion sections, where there is an association of shared experience as well as reporting past research in both sections. In their analysis of 15 medical research articles Nwogu and Bloor (1991) found that introduction and discussion sections have simple thematic structures (associated with explanation and argumentation) while methods and results sections have relatively more constant theme structures (associated with

description).

3.44 Methods and Results sections in research articles.

In most cases, especially in structural chemistry, the methods section (incorporating materials and methods sections, and experimental sections) is the linear version of the lab-book, a "listing of procedural formulae" (Swales 1990:121) with details of techniques, brand names involved in techniques, temperatures, measures of amounts used, reaction speed, molecular size (mml, mhz, mmol, respectively) and so on. Swales states that these sections are "highly abstracted reformulations of final outcomes in which an enormous amount is taken for granted" (1990:121). Certainly, these sections are the most inaccessible to the non-initiate, although for the intended user they may constitute the first port of call in terms of the indexical use of language we mentioned earlier. Swales' point is that this seems to belie the empirical ideal that explicit detail ensures the theoretical possibility of replicability.

Methods sections are characterised in the literature by the agentless passive and Hania and Akhtar (1985) find that the use of the passive exceeds the use of the active (54% to 46%) in the methods sections of MSc theses in the experimental sciences. The passive is commonly said to enable a distancing of responsibility of actions from the actual protagonists (Heslot 1982, Swales 1990:120). Sager et al. (1980:209) note that the use of the passive in technical writing is a result of the need to thematicise the result of an action giving informational weight to the action (expressed by the verb) that is ascribed as the cause.

According to Swales (1990:121) both methods and results sections are "mutually inter-dependent". The literature points to formal similarities between these sections. Adams-Smith (1984) has analysed authorial comment (in terms of modality, first person pronouns, markers of analogy (*like*) and use of discourse items such as *possible*) and found that the distribution throughout IMRD sections decreases slightly in the methods and results section and increases slightly in the discussion section. She also finds that the use of the past and the passive follows this pattern, except that they increase in the MR sections. West (1980) has also demonstrated that *that*-nominalisation is extremely rare in methods and results sections, while frequent in introduction and discussion sections. This is corraborated by Brett (1944) 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 continuous/progressive) tenses being rare.

3.45 Discussion sections in research articles.

Understandably few studies have looked at discussion sessions alone, and comparative studies have emphasised the similarity of grammatical features with introduction sections (Guntzman and Oldenburg 1992). It could be argued that the rhetorical functions of discussion sections are very different, since they provide the synthesis of results and their evaluation as viable elements of a new model. Swales (1990) has suggested that discussion sections are something like the mirror images of introduction sections seen as looking out from the research into the wider world: thus introductions synthesise past research and evaluate old models inwards, while the discussion section does the reverse. This does not explain why grammatical features are shared, although it seems to suggest that the surface items studied so far do not account for rhetorical perspective.

Hopkins and Dudley-Evans (1988:117) state that discussion sections have one obligatory move of Swales' original 11 (1981a): 'statement of result' and that discussion sections follow a cycle of moves involving a statement outlining variable *n*, description of previous research relating to variable *n*, evaluation of this research (optional), statement outlining variable *n+1* and so on. Adams Smith's (1984) analysis of various forms of authorial comment (evaluation in adverbs such as *insufficiently*, adjectives *important, minor*, verbs *established, claimed* reporting nouns *speculation, hypothesis*) qualifies the discussion section in the *British Medical Journal* as the most subjective section of the research article. Hanania and Akhtar have characterised the modal as a typical form in the discussion section, associating it with making "qualified generalisations" (1985:53).

3.5 Authority and science writing: Myers

Objective fact is only what the dominating group thinks it is. (Kaplan and Grabe 1992:200)

In describing the constraints on free variation in scientific texts in the previous discussion of the research article, it can be seen that convention plays a part in delimiting the extent of permissible variation. These variations and constraints, for example in terms of verb tense or modality, are inherently concerned with decisions of theoretical acceptability imposed by authority.

In their views on authority in the scientific community, Kaplan and Grabe (1992) represent

what we might call the 'stormtroopers' of constructivism. They reject the commonly held view of technical writing experts that scientific writing has specific forms because of its attempts to present neutral, replicable and falsifiable facts. Instead they claim that all written texts reflect unconscious 'rhetorical assumptions' about the format, rhetorical structure, and linguistic organisation of text that preclude transparency. They further claim that this obscuration enables the scientific discourse community to police itself and to establish a hierarchy where access to discourse is limited according to the concerns of maintenance of coherence of the discipline.

In his procedural study of the rejection, editing and eventual publishing of five scientific articles in molecular microbiology, Myers (1990) sees science as a social process that produces scientific knowledge as text. He particularly criticises linguistic and ethnographic studies that see text as an empty vessel in opposition to practice, and finite fixed science opposed to the 'real business' of social activity. Text for Myers is a form of a conceptual knowledge structure and a social consensus; scientific reality is not transformed but rather formed by text:

> Content of natural knowledge cannot be separated from the social processes that produce it. (Myers 1990:20)

He traces the social and literary studies of science from the purely institutional analysis that revealed little about how authority is established or how the knowledge structure is rewritten. Myers exploits the circularity of scientific observation: only a rigorous experiment can show the true nature of the phenomenon, but the rigour of the experiment can only be judged on whether it reveals the true nature of the phenomenon. Since this is circular, consensus about scientific knowledge must be negotiated socially. One way in which this is done is by changing the level of claims (from originality to following on from previous work) where decisions about where the research fits into a new discipline are made in the conceptual framework and more specifically in the terminology chosen and the terminological changes proposed by the reviewer. Myers cites one author's choice of *reproductive processes* instead of *reproductive behaviour* to fit in better with the new field of physiology (1990: 52). At a more rhetorical level, the separate presentation of a hypothesis and data indicates to Myers that a writer had not evolved the 'sufficient syntax' to connect the two conceptually (1990:54). From changes writers use to weaken their claims, Myers builds up a 'rhetoric of assertion' where distancing from a certain position is achieved by placement of data, the use of the passive, long noun phrases and hedging verbs. In general, the negotiation of claims that takes place over the period of rewriting

depends on issues of future lines of research, appropriateness of claims to the journal, organisation and length of claims, the position of the researcher in the hierarchy of the discipline and the use of literature to back up data or to support claims (1990:99). Myers finds that the observation of the writing processes of science should start from a non-ironic basis of rhetoric which is pervasive in all aspects of the production of text. He also finds that the social context needs to be reconstructed, including paying attention to texts related to the production of the target text as well as the role of authority in the reformulation of these texts. Finally Myers finds that all forms of discourse share fundamental principles and all have a role to play in the social construction and renegotiation of knowledge:

> Though scientific texts come out of an unusual social structure, and thus are different in some details from texts in other discourses, they are not doing something fundamentally different from other texts... Science uses our language and despite attempts to purify it, it is still loaded with social and political implications. (1990:258)

With the constructivist stance on science in mind, we can now turn to models of discourse analysis and functional grammar that are compatible with this approach.

CHAPTER FOUR: REFORMULATION IN SCIENTIFIC DISCOURSE.

4.0 Reformulation.

In the previous section, a general characterisation of the research article was proposed in terms of linguistic analysis of a generally functional nature. In this section, linguistic methodologies of text analysis are explored in more detail to provide a framework in order to study to complement those studies of scientific texts that have already been elaborated. The functional framework will prove particularly relevant to the previous discussion of rhetoric and authority in scientific writing, and once the theoretical linguistic issues have been aired here, methods of their implementation in the computerised analysis of large corpora are outlined in the following chapter.

4.1 Discourse Analysis

> You can argue with a claim, but you can't argue with a nominal group. (Halliday and Martin 1993:39)

Halliday's view of scientific language can be encapsulated as a practice that constructs our world view as opposed to both the internal (cognitive, mentalist or psychological) and the external (sociological, ethno-cultural) approaches which see language as a non-determining reflection of mental processes or social context. Discourse therefore becomes the fundamental unit of language at a social-semiotic level. A text is bound therefore to be a discourse, it cannot be disassociated from its context (as in Chomskyan formal grammars) and cannot be considered to be a complex grammatical realisation of another text or set of propositions (as in the textlinguistics of de Beaugrande and Dressler 1981:89).

Halliday and Martin (1993) see scientific discourse as part of the authoritative system of hierarchical control in a similar way to Myers (1990). They have drawn attention to the pervasive effects of scientific linguistic procedures on our everyday language and to the alienating effect of scientific language on those who have not been trained to handle such discourse. This is not the same as incomprehension, but the feeling that a language is being used in an exclusive way. They say of scientific English:

> It 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. (1993:4)

According therefore to Firth's polysystemic approach, when a society changes its system of self-expression, its original resources of expression are adapted to a new framework taking on new roles (Halliday and Martin 1993:9). The evidence for this lies in the fact that there have been (according to Halliday) shifts in the use of grammatical metaphor on a major scale which correspond to political, social but above all technological upheaval, as in ancient Greek, mediaeval Latin and renaissance English. Halliday and Martin argue that the same processes are still evolving in modern scientific English. In English, the refinement and change of role of the nominal group have meant that information can be reformulated with greater flexibility within the clause. The problem with this is that as nominal groups compact ideas and propositions, so they become increasingly difficult to interpret separately. This may have something to do with the terminological resources of complex nominals and nominal clauses in that once formed they become idiomatic and to some extent beyond interpretation on the basis of the individual elements.

4.2 Choice in a probabilistic grammar: Firth and Halliday.

Halliday's systemic grammar has become a influential account of language for researchers concerned with language as a social system. As mentioned earlier, both Halliday and Sinclair were students of Firth, and Halliday views their work as essentially having a common goal: an approach to 'wording' or the patterns of reformulation in language that are realised by variation in the lexicogrammar (1992:63). Unlike the heavily structural accounts of transformational and generative semantic grammars, systemic grammar has very little 'hidden' or deep structure, in that it does not attempt to classify each instance of linguistic variation in terms of further branches or transformations in a tree-structure or in terms of more semantically basic categories.

Halliday (1985) states that systemic grammar is essentially semantic, and his analysis depends on his early elaboration of three metafunctions, mentioned above in their relation to Sager et al's (1980) model of textual production. These evolved from Halliday's work on intonation and the possibilities of expressing variable emphasis of mood and theme in spoken English. The transitive or ideational function is an expression of the psychological representation of reality perceived as participants, processes and circumstances in language

(Halliday and Hasan 1989:68), and can be seen in the distinction between the grammatical subject (*That teapot*) and the psychological subject (*the Duke*) in Halliday's famous example: *That teapot was given to my Aunt by the Duke.* The thematic or textual function mediates the way the message is presented in discourse, such that *My Aunt* could be placed in prominent thematic position as the main element of informational background in which the action of the message is to take place. Finally the modal or interpersonal function (mood) expresses the message as an arguable proposal (Subject and Finite) together with Predicator, Complement and Adjunct.

The metafunctions present an overall organisation to Halliday' grammar which sees each area of grammatical variation as a choice point at a certain linguistic level. At the level of the clause, the choice may be realised:

- in terms of complex or simple mood (indicative: declarative/interrogative ; imperative, infinitive, participle/ gerund)
- in terms of transitivity (material/mental or relational process; middle 'ergative' or effective 'active/beneficiary' participants; time/cause circumstantials),
- in terms of theme (unmarked or marked)
- in terms of polarity (positive or negative)

At the level of the verb group choice may be realised:

- in terms of deixis (primary tense as past/present/future or some form of modality)
- in terms of secondary time (none, 1 or 2); phased [began to take] or unphased [took])
- in terms of voice (active or passive)

The above system is clearly not complete (as adapted from Halliday 1992:67) but indicates the increasing delicacy of the system as units become smaller. It is also important to note that since each choice mediates others along the line, once a choice such as transitivity = 'material process' has been made, this delimits choice available about types of verbs, tenses involved as well as participants at the same level of choice. Some choices are also obligatory corollaries of others, so that once a choice to include secondary tense has been made, a choice of phased or unphased (continuous or simple) aspect must also be made.

Halliday (1991,1992) has proposed that choices in functional grammar operate on a largely probabilistic basis, as Sinclair states : "in many binary systems the frequency of one choice seems to occur roughly one order of magnitude more commonly than the other." (1993c:167). Thus high redundancy would give equal probabilities to all members of the systemic choice, for instance the probability of present/past or future simple tenses may be 33%. Whereas low redundancy would assign a greater probability to the least marked

choice, so for example positive polarity should in theory receive a 90% probability and negative polarity 10%. Halliday and James (1993) have established from a very basic statistical analysis of 25 high frequency verbs in the 20 million word Cobuild corpus that polarity and primary tense do appear to be distributed in this way. This clearly has implications for the automatic analysis of the language, since as Halliday says: "frequency in the corpus is the instantiation (note, not realization) of probability in the grammar." (1992:66). One example of this, Barber (1962) calculated that of 1770 verbs observed in astronomy, biochemistry and electronics 89% (65% in the active voice) are of the simple present and 11% of other tenses.

Halliday also points out that there may be increasing complexity in the system as it functions in running text. In scientific and technical discourse, this complexity may be expressed as the marked effect of grammatical metaphor in the 'logogenetic history' of a text where the original verbal process is transformed into the Actor and then evolves by nominalization with increasing numbers of properties (adapted from 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.
```

Grammatical metaphor is clearly an essential linguistic tool in the description of reformulation, and is set out in more detail in the Data analysis (Chapter 10). What can be retained here is a principle of reformulation based on Halliday's observations of choice within a linguistic system. *How glass cracks* and *glass fracture growth rate* are both very different propositions, but they have evolved within a running discourse through several reformulations. If reformulation is considered to be the linguistic rearrangement of the same concepts such that their syntactic status is different then we can contrast this to the collocational idea of recontextualisation (similar to Pavel's virus-like *LSP-collocations*) as the gradual evolution juxtaposing one proposition with another within a new co-text. Both processes fundamentally involve semantic reconceptualisation in the long run. Let us look at more detail at how discourse analysis has tackled the question of recontextualisation in and long range collocation.

4.3 Lexis and recontextualisation: Hoey.

Texts have a rhetorical structure that is hard to define in terms of explicit linguistic items, or even in terms of labels that could be applied from one utterance to another. In a key study in discourse analysis, Halliday and Hasan established that there are a series of explicit

relations that distinguish a text from a string of sentences (1976:6-7), provided that active interpretation by the participants can be demonstrated: "Cohesion occurs where the *interpretation* of some element in the discourse is dependent on that of another." (1976:4).

In the cohesion model, significant areas of what would otherwise be considered sentence grammar are involved in co-reference: signalling links either outside the text (exophora) or backwards and forwards beyond the level of the sentence (endophora). Reference, the first category, involves demonstratives, pronouns and comparatives. The second and third categories are substitution and ellipsis: demonstrative or syntactic reformulations of nominal and verbal groups or clauses. Conjunction constitutes an intermediate category of reference that includes not only conjunctions, but also idiomatised clauses (*that is*) and lexical items such as adverbs (*previously*) (1976:242). Finally, lexical cohesion involves reiteration involving exact repetition, synonyms or superordinate words (1976:278) and collocation, the occurrence of lexical items that "share the same lexical environment." (1976:286).

Hasan (1984) and Hoey (1991a) have claimed that lexical cohesion, is of much greater importance and complexity than the other categories. Hoey (1991a:9) calculates that Halliday and Hasan's own analysis of seven texts classifies 42% of instances as lexical cohesion while grammatical reference accounts for 32%. Halliday and Hasan's view of collocation is also problematic in that it covers all other semantic relations that are not direct reference yet involves examples of lexically related or associated items. Hoey argues that if cohesion is organised lexically, sentences which share many lexical referents can be interpreted as a coherent whole. In fact, if there are multiple links between sentences, these sentences could be interpreted together as a coherent text (1991a:192) (i.e. as a unity in Halliday and Hasan's terms). Importantly, Hoey emphasises that bonded sentences bring with them a coherence that is more than the cohesive elements that brought them into relation in the first place, we begin to see a similarity here with Pavel's 'terminology in the making'. Thus if all the bonded sentences are then presented as a whole, they may turn out to be a valid summary of the original text (1991a:34).

The idea of long-range collocational structure in text was established by Phillips (1985) whose idea of 'aboutness' is expressed by the statistical inter-collocation of collocates (1985:100). Phillips starts off from the basis of the purely statistical analysis of lexical collocation in the chapters of a technical text book. After eliminating all grammatical words from his analysis, all lexical pairs that co-occur significantly were then tested by cluster analysis to see whether they collocate significantly with other highly collocating words

(1985:86-87). The result is that collocation is text sensitive: the networks of collocating words can be seen to change from text to text indicating the basic topic structure of the text. Phillips also established the rule of thumb of a minimum of 3 links in one sentence to link chapters and therefore repetition aids long-distance organisation of text. Källgren (1988a, 1988b) has used a similar methodology to obtain automatic abstracts of texts on the basis of intercollocation. Broek and Trabasso (1986), working on causal hierarchies in textgrammar, and Alterman and Bookman (1990), measuring connected paths in story grammars, have also found that it is the nodes with the most connections (or synonymic/taxonomic matches) in networks rather than those which are hierarchically or grammatically prominent are generally chosen for inclusion in summaries.

Since any sentence of this discourse can be related to another, the role of lexical cohesion is to allow the interpreter enough signalling to make relevant links in what Hoey terms a network (1983:176). If signalling of all types aids the formation of a network, Hoey predicts that discourse is "non-linearly organised" (1983:177) rather than set out in an entirely implicit dialogue between signaller and interpreter. The role of lexical cohesion is therefore to reformulate established concepts in the light of new ones:

> There is informational value to repetition, in that it provides a framework for interpreting what is changed. (Hoey 1991a:20)

This is a concept referred to as instantial meaning by Halliday and Hasan (1976:289) and established very early on in structuralist terms:

> ...la valeur d'un terme peut être modifiée sans qu'on touche ni à son sens ni à ses sons, mais seulement par le fait que tel autre terme voisin aura subi une modification. (De Saussure, 1916:167)

This observation, Hoey claims, is backed up by analysis of close testing where informants can guess as the meanings of unfamiliar words in context: thus Hoey's conclusion is that language is stored as a whole very much as received (1991a:154, and this may explain how meanings are acquired and how collocations are recognised over long distances (1991a:155). As he says,

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

Hoey has introduced a clear framework within which lexical cohesion can be linked to our understanding of text. Lexical organization can in turn can be contrasted with the function

of the text and the ongoing context as it reinforces, modifies and creates collocations (1991a:210):

> Every lexical selection affects or creates cohesive links that [...] help organize the text; patterns of organization of a more conventional kind, such as a problem-solution patterning, likewise only have reality in so far as they are made by lexis. Conversely, relations between lexical items, for example, sets, collocations, are a function of their appearance in the text. Furthermore, each textual selection constrains the lexical choices possible, and it is in the combination of the lexical and textual choices that [the writers' or speakers'] creativity is expressed. (Hoey 1991a:217)

Collocation in a lexical network is clearly not the same as lexical repetition, for which we might use the term lexical chains (Halliday and Hasan 1989). Lundquist (1989) has argued that non-experts reading scientific text rely on lexical networks to establish long-range links, and that experts do not need explicit signalling and are thus able to skip and skim text and establish global relations. For example, only an expert knowing what *aprotinine* does and is would be able to guess the implications of *aprotinine reacts favourably with...* where there are no surface clues and knowledge of measurement and tools necessary for scientific experimentation is necessary to classify the evaluation as a meaningful argument. (1989:141). Coherence is therefore dependent on background knowledge, whatever the lexical cohesive network tells us. This argument is confirmed by Myers (1991:13) whose analysis of cohesion in scientific articles reveals the complexity involved in deciding the connectedness of lexical repetitions especially in terms of synonyms (*DNA - genome*) superordinates (*molecule-product of transcription*). Myers (1991:5) argues that background knowledge of the scientific paradigm is essential for any networks to be built up, and this would account for the differing forms of cohesive devices used in scientific and popularised texts. He also 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)

Following Källgren, Hoey posits then that 3 lexical links between any two sentences are enough evidence to show that the sentences form a 'lexical bond' (1991a:125). Hoey states that only repetition of old information counts (1991a:69): thus *scientists* is not linked if it is followed by *biologists* because new information is being introduced by *biologists*. However, superordinates do count: *bears* would be linked to *animals*. Unwanted cohesion, as Hoey calls it (1991a:46) includes abbreviated pronouns and determiners which do not, he claims, go beyond intersentential relations. Simple repetition consists of exact repetition

of open set items (1991a:167). Complex lexical repetition includes items which share lexical morphemes nut are not formally identical (*happy* | *unhappy*) or items that are formally identical, but have different grammatical functions (*drug* | *drugging*). At the next level of lexical cohesion, Hoey identifies simple (*statesman* | *politician*) and complex paraphrase (*hot* | *cold*) which may also occur as a result of the triangular link: thus *author* _ *writer* (simple paraphrase) _ *writings* (complex repetition). These formal or semantic links preclude collocations such as *christmas ~ carol* and *sickness ~ doctor.*

Halliday and Hasan recognise the claim that reference is not essential for a "cohesive force" to be set up between lexical items. For example, a repeated referent can be non-inclusive (1976:283):

#1 There's a boy climbing that tree.
#2 Most boys love climbing trees.

Hoey (1991a) has pointed out that examples like "*wet...dry, sky...sunshine, order... obey*" are not true collocates in the Firthian sense of *textually* co-occurring lexical items. Instead Hoey (1991a) has attempted to relate large scale non-linear patterns of cohesion to coherence. Hoey bases his argument on the prevalence of lexical cohesion and on Winter's (1978) work on the signalling function of three types of lexis (subordinators, sentence connectors and lexical items). His initial position is that coherence depends on an underlying problem-solution structure in discourse. For example, most fluent speakers of English can rearrange randomly mixed sentences into a coherent order, as in Hoey's sentry text:

SITUATION	I was on sentry duty.
PROBLEM	I saw the enemy approaching.
SOLUTION	I opened fire.
EVALUATION	I beat off the enemy attack. (1977:12)

Hoey's claim is that there is a consensus in the way that we approach discourse that expects a problem-solution structure unless clues from the text indicate otherwise. We can see that cohesion is operating in two areas in the sentry text: in terms of syntactic equivalence (Hoey 1991b:397) similar to Firth's *colligation* and Halliday's *grammatical metaphor* (*thematic first person subject + past tense predicate*), and lexical cohesion (the military semantic field: *sentry duty, enemy attack, open fire*). Hoey's point is that coherence can be seen as a structure we impose on discourse, while cohesion is more to do with modifying or reorganising our initial coherence predictions on the basis of tangible linguistic evidence. It is possible to rearrange the sentry text again with explicit signalling: *I beat off the enemy*

attack because I opened fire when I saw the enemy approaching. Thus cohesion, of which lexis is the most complex of forms, is seen as textual organisation, which Hoey is keen to distinguish from structure.

There is therefore considerable theoretical support for a notion of recontextualisation in terms of the organisation of lexical cohesion in text. A problem still remains however concerning the relation of lexical cohesion and coherence. Since Sinclair's views on corpus linguistics are to make up a considerable methodolohy in our data analysis, it is necessary to take account of his theory that attempts to relate a notion of textuality or coherence to aspects of lexical signalling and phraseology.

4.4 Postures and planes of discourse: Sinclair.

Sinclair's view of discourse can be seen as a dynamic series of participant-expressed positions, or postures, each one being the basis for an interpretation of what has preceded and what might follow. A posture, states Sinclair is a "unit of discoursal meaning or behaviour." and a "logical consistency change in attitude or circumstance." (1993b). Since postures are either maintained or changed in spoken as well as written discourse, maintenance and change provides discourse with a primary structure for analysis.

Sinclair's position shifts the emphasis of discourse analysis and text linguistics from an incremental bit-by-bit view of coherence, where coherence is seen to be the sum of the whole of the discourse, to one where understanding is brought from one complete state to another utterance by utterance: "a text is represented at any moment of interpretation by a single sentence" (1993d:7). Out of context, a sentence like *We begin our programme on 9 July* is simply informative, but placed in front of the request *Can we have an official response from you regarding these suggestions?* the sentence takes on an implicit persuasive function (1993d:6). The reader infers coherence about what the text has been and will be about on the basis of clues in the sentence in question (1980: 253). Yet the exact nature of these clues varies: they may be implicitly inferred from semantic world-knowledge, or they may be built on explicit cohesive structures as described by Halliday and Hasan (1976). Sinclair claims that the way participants make interactive inferences about preceding and following discourse depends on the explicit clues which include logical connectors (*therefore, so, however*) and items in the spoken language such as *anyway, you see, I mean*. These make up what Sinclair terms the "interactive apparatus of the language." (1993d:7).

Sinclair's interactive view of discourse precludes the conscious building up of 'point-to-point' cohesive networks, either in the spoken or the written language. Only the latest utterance is recoverable in its original linguistic form, while the complex structures of coreferents and themes in the written language are only present non-linguistically in the mind of the language user (1993b). Thus encapsulation of past discourse and predicting of future discourse is essential for the processing of language and has been suggested by Winter's (1977) cataphoric 'anticipation' and anaphoric 'retrospective' systems of reference. Sinclair characterises any prediction of forthcoming discourse as prospection and the backward interpretation of discourse as retrospection (1980, 1981) later termed encapsulation (1993b,1993d:8). Sinclair finds in his exploratory study of one expository text, that encapsulation accounts for the majority of coherent sentence links, and claims that where there are no explicit links encapsulation will be the default interpretation (1993d:22). Such links as there are may be in terms of logical connectors or deictic reference to specific referents, a process he terms refocussing (1993b). In addition, links may be established by rephrasing (1993b) by 'verbal echoes' such as *perceived disadvantage...perceiving itself to be a disadvantage* or paraphrases termed 'overlays' such as *by studying those of their rivals.... to keep in touch with trends in other countries* (1993d:17).

Prospection appears to be less prevalent than encapsulation, but Sinclair claims that it constitutes the primary explicit structural feature of the text:

> Prospective structures are concerned with control over what happens next. They are the attempts of one individual to pre-classify the next utterance of another, to negotiate speaking rights or to indicate desire or willingness to relinquish them. They are understood with reference to a finite network of possibilities, organized in a hierarchy of units. (1980:254)

The mechanics of prospection have been explored by Tadros (1985: discussed below), although Sinclair discusses two categories termed 'attribution' which involves some pre-classification of discourse (*his message..., to quote The Prince of Wales..., the statement...*) and a version of Tadros's 'advance labelling' (*The implications..., a flexible response..., The notion of perceived disadvantage...*) where the concept is elucidated in the following sentence. Prospected sentences cannot therefore encapsulate their prospecting sentences, they either prospect or are encapsulated by the following sentence.

This view of 'stance' in discourse can be seen as an extension of Sinclair and Coulthard's (1975) labelling of interactive rhetorical moves in spoken discourse. The process of

encapsulation in turn constitutes the theoretical notion of topic or 'aboutness' (Phillips 1985) of text based on long-range semantic or cohesive links, where the latest utterance or sentence constitutes an encapsulation of the discourse that has preceded. Sinclair expects links to be worked out on an *ad hoc* basis, relying less on explicit thematic networks than on a sense of prosody and phraseology, an idea later followed up by Louw (1993):

> Some recall links may not be very strong, but may be brought into focus by a clearer pattern nearby - like sound patterns in poetry. The reader or listener is often aware merely of a semantic coherence running through the discourse, which can be named at any time as the topic or theme. (Sinclair 1980:255)

As with this directional view of discourse in terms of backwards and forwards construction of coherence, Sinclair formulates a procedure for explaining how written text is just as interactive as spoken text:

> Language in use has two aspects: at one end and the same time it is both a continuous negotiation between participants, and a developing record of experience. (Sinclair 1981:2).

Sinclair thus sees topic in language as the construct of two planes. The autonomous plane linguistically internalises, organises and updates the topic, making a record of experience which may be contrasted with new material and relies on past shared experience (1981:4-5). The interactive plane develops the topic on the basis of a perceived interaction of the participants (1981:2-4). Sinclair argues that both spoken and written utterances can be describable in terms of both planes, since both have equivalent mechanisms such as written language's ability to interact by signalling predictions (Tadros 1985) or by hedging or failing to signal the real participant ('It is interesting to note that...' (1981:6)).

Sinclair states that where prospection and encapsulation break down, there is a plane change (1981:8, 1993d:8). In the sentence *David Blunkett knew that.* 'that' encapsulates explicitly in the interactive plane a complex series of propositions that may have been implicitly built up over time (i.e. in the autonomous plane). Yet if *David Blunkett* has never been referred to before, there is a change of posture and a plane change: we are now expecting some interactive statement to justify his presence in the autonomous plane. Uncooperative counter-questions (*What did you just say?*), on the other hand, can be seen as encoding the dialogue exclusively in the autonomous plane by explicitly referring to the interaction, deviating away from a previously implicitly interactive dialogue. The autonomous and interpersonal planes can thus be seen as the linguistic and social systems of cohesion (either encoding or negotiating propositions) operating in the background of

prospective strategies that are employed to establish coherence. The maintenance and change of postures could account for the extended structure of the written language. Sinclair claims that independent clauses have their own posture whereas subordination allows for the same posture and proposition to be continued without any plane change. Similarly, nominalisation can be seen as "protecting a proposition from a truth value" (1993b). If structures can be embedded within the same posture, this gives the written sentence a role that is just as interactive as the spoken utterance.

In the context of Sinclair's interactive view of spoken and written language Tadros (1985) elaborates a theory of prediction in text, where the reader may either be explicitly told what is going to happen i.e. prediction, or the reader may infer what is going to happen, a process Tadros terms anticipation (1985:6). She identifies six general categories of prediction:

Type 1: where the writer promises to enumerate a series of points. (*as follows:, 3 advantages, Firstly*) (1985:14-22)
Type 2: where the writer labels his/her discourse in advance (*As Pigiou defines.., c.f. table 1, why does X overlap with Y?*) (1985:22-28)
Type 3: where the writer detaches him/herself from what others say by reporting (*In their view, argue/say/emphasizes...that*) (1985:28-35)
Type 4: where the writer recapitulates the main points of her/his own text (*as mentioned earlier, so far considered, make reference*) (1985: 35-42)
Type 5: where the writer detaches him/herself hypothetically (*suppose, let X be...*) (1985:42-48)
Type 6: where the writer allows for prediction on the basis of a rhetorical question (1985:49-52)

Such forms of prediction have particular phraseological characteristics. For example, many 'sub-technical' nouns appear in cases of enumeration (*advantages, conditions, stages*) (1985:14), while verbs involved with reporting generally take a 'that' complement (1985:31) and in cases of recapitulation there is usually an initial circumstantial element (*In previous chapters, In summary, In this study*) (1985:40). Although Tadros emphasizes the fact that prediction can only occur in structures of predictive and predicted utterances, in some instances it is hard to tell whether a particular element is the cause of prediction or whether the utterance as a whole has to be taken into account. For example in the following predictive-predicted pair (#1) only the Vocabulary 3 element (Winter 1977): *distinguish* is picked out by Tadros as a predictive act of distinguishing two terms, yet instinctively we may be able to argue that the distinction is set up by the phrase *It is important to..*:

#1- *It is important to distinguish between real and nominal wages. Nominal wages are in terms of money...Wages are wanted only for what they will buy.* (1985:25)

If we take into account Sinclair's view of an prospective utterance providing clues as a whole, then segmentation into smaller units may not be such a problem. The essential point is that the entire text can be seen as functioning in terms of interaction not just as the linear laying down of facts, as Tadros concludes:

> ..there are major rhetorical organizational features which go beyond presenting propositional content in a suitable way. There is evidence that the writer does not simply present facts and ideas to the reader, but is rather concerned that these should be understood and accepted. (1985:63)

Hazadiah (1993) has similarly taken up Sinclair's idea of prospection to argue that topic occupies a higher discourse rank than the sentence. Hazadiah argues that if spoken discourse is often most natural when it is implicit (1993:57) then the topic can be seen not as a product that is 'agreed' early on, to be developed by the ensuing discourse, but as a process that emerges from the gradual building up of a conversation. The concept of prospection, then, is used to describe the "presupposition of support" (1993:57) that is based on the current spoken exchange. The presupposition is either fulfilled or the topic is discarded and this constitutes the dynamic that allows the topic to be built up over several exchanges. Thus at each point in the discourse there is a "topic potential" (1993:60). If the topic is unsupported in the interaction, for example if one participant evades a question, then there may be a "plane change" (1993:61) as outlined by Sinclair (1981), where the interactive potential is unrealised and recorded in the autonomous plane as some kind of metadiscoursal event, such as 'evasion'.

A formal model of relations between sentences has been proposed by Mann and Thompson (1986, 1988). In Rhetorical Structure Theory (RST), each sentence is connected to the next by a hidden rhetorical relation, the most basic being *circumstance, contrast, joint, forwards motivation, backwards enablement, sequence*. Since the relationship is hidden, there may be no lexical clues to the relationship, but the theory provides logical constraints for each. The relation 'evidence' therefore depends on a constraint on the initial sentence (the 'node' in the case of evidence) that readers might not believe it, and on the following sentence (the 'satellite' in this case) that it might be found credible. The idea of nuclearity exploits the intuition that certain clauses are independent, essential to the purposes of the author and therefore cannot be substituted as easily as others in the discourse. These relations may stretch across several sentences, leaving a skeletal structure of node sentences or clauses. Mann and Thompson (1988:267) claim that it is possible to form a coherent 'synopsis' from these. Moore and Pollach (1992) have criticised the linear properties of RST arguing

that relations between elements of discourse take place simultaneously on other discourse levels. In the following sentences, an 'evidence' relation may be held simultaneously as a 'volitional cause' relation:

1- George Bush supports big business.
2- He's sure to veto House Bill 1711.

If simultaneous relations holding between sentences depend on the perspective of the reader, this would appear to support Sinclair's view that coherence is a largely a matter for individual choice in context. This also raises the point made in Chapter 3 on the use of abstracts, that coherence may be inferred from a partial reading of the text, or a non-linear reading of the text. This might indicate that, for the reader who stands the greatest chance of inferring coherent relations, a non-linear approach to the text could successfully circumvent such devices as thematic progression.

4.5 Linear theme versus non-linear prospection

Mauranen (1993) demonstrates the important role of sentence themes in the establishment of prospection. She proposes that if there is insufficient semantic material in the ideational theme to refer back to the preceding themes then some kind of "orienting theme" is necessary to help the interaction along (1993:96). This either precedes the theme in the following sentence, or is somehow signalled in the preceding sentence. However, Daneš's view is that theme picks out the most relevant point of information from a mass of previously built up information. Mauranen argues that if we assume that the propositional content is built up over time, as in encapsulation, then selection of theme also has to affect what is going to follow (1993:102). The actual criteria for what constitutes theme, Mauranen states, has been established by reference chains and lexical cohesion (as in Hoey 1991a and Källgren 1979). Mauranen finds that there is often no clear lexical or referential link between two sentences, and that there must therefore be varying scope for prospection. She gives the following three-sentence example from a biology text (Mauranen 1993:111) [I have emphasized the sentence themes]:

#1 What mediates the increase in platelet activity in pregnancy is unknown.

#2 Formation of thromboxane A2, the major cylcoxygenase product of arachidonic acid in platelets and a potent platelet aggregant, is reported to be increased in pregnancy.

#3 A possible source of increased thromboxane formation in pregnancy is the placenta, which has been shown to generate thromboxane in vitro.

Here Mauranen sees the theme in sentence #1 as a firm but unspecific prospection, where the theme in #2 specifies the topic without having to establish its own relevance. But in #3 the thematic relevance has to be re-established because #2 has not prospected it. Here, #2 is encapsulated by #3 in a nominalisation of #2's main propositional content within #3's theme. It is this type of theme that Mauranen identifies as an 'orienting theme.' (1993:112).

Wikberg has attempted to establish the relationship between Hallidayan thematic structure, Hoey's and Källgren's lexical cohesion and her own view of discourse topic, which appears as an umbrella concept accommodating theme and cohesion: "how a given subject matter is manifested lexically" (1990:232). Thematic structure is differentiated from lexical cohesion by its dynamic relationship and the management of given and new information in a discourse. Lexical cohesion represents thematic links but also contributes, by effects of recurrence, to "building a textual world" (1990:231). Comparing head words in thematic and rhematic positions in 4 chapters from 3 expository books, Wikberg finds that as the span between one lexical item and its previous referent expands, the difference between thematic and rhematic theme becomes less interesting: "hierarchical structural features take over instead as the dominating influence across paragraph boundaries." (1990:246). The rigid macrostructure of topic, particularly of texts that are non-narrative, is therefore more important than thematic progression.

Conversely, Nwogu and Bloor (1991) have found that subject matter in the form of lexico-semantic networks does not have as much bearing on the organisation of information in texts as do functional and contextual constraints. Comparing thematic progressions in 15 medical research articles and their corresponding abstracts and journalistic accounts, they find that journalistic accounts tend to have simple linear themes (where a new theme leads on from a rheme) and these tend to correspond to paragraph constructions based on explanation and argumentation. Research articles tend to have more constant themes (where theme stays the same) and argumentative and explanatory paragraphs tend to occur in the introduction and discussion sections. The themes in research articles and abstracts also tend to be ideational themes, often adjuncts introducing passive clauses, while those of journalistic articles tend to be textual and interpersonal themes (proper nouns and personal pronouns). Abstracts make equal use of both types of thematic progression but do not have any complex split rhemes or derived themes, the kinds of long-range structure found over a span of three or more sentences that would be inappropriate for the genre.

CHAPTER FIVE: PHRASEOLOGY AND CORPUS LINGUISTICS.

5.1 Corpus Linguistics: The automatic analysis of texts.

> 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, 1953, para.199)

The corpus approach is of central interest to this thesis, firstly because specific genres have only very rarely been scrutinised by computers and secondly, the computational analysis of a genre promises to be a particularly fruitful tool given the essentially context-based approach which has been outlined above. On a more specific note, the finding that genre analysis has produced little research on general lexical analysis of scientific abstracts and the growing interest in the role of lexis and phraseology in discourse is a particularly attractive opportunity, given that these are exactly the areas that have been most fruitful in corpus linguistics.

5.2 Corpus Linguistics and the description of language.

One characterisation of corpus linguistics, although admittedly a narrow one, is the quantitative automatic collating of linguistic features from a computer held reference corpus representative of some part of the language (from many samples to a specific genre). The use of computers for data collection has meant an increase in corpus size as well as refinement in the types of data that can be automatically collated. 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) has identified several common currents in corpus work and has baptized the field *Computational Corpus Linguistics* (CCL). He sees CCL as a research program as opposed to a paradigm, although elements that unite workers in the field are an interest in the empirical, quantitative description of performance related aspects of language as opposed to Chomsky and his followers' search for universal rational rules of competence. According to Leech CCL practitioners regard authentic data and pattern-searching as essential methodological goals. The main advantage of this 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' in order to quantitatively test the fit of models worked out on other corpora. A corpus-based model of linguistic behaviour is therefore falsifiable because it can be tested against fresh data.

Computational corpus linguistics has been concerned with 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 (Papegaaij and Schubert 1986), the development of terminological knowledge banks (Ahmad et al. 1990, 1991) and the development of language teaching materials and programmes (Willis 1990, Johns and King 1993).

5.3 Developments in corpus linguistics.

The Brown corpus (Kučera and Francis 1967) of 1 million words was the first electronic store of texts for systematic linguistic analysis of discourse with the underlying aim to be as representative of the general language as possible. The London-Oslo-Bergen (LOB) corpus of 1 million (originally 350 000) words (Svartvik and Quirk 1980, Leech 1987) contains 15 types of written text (maximum length 2 000 words), and constitutes a major source of data for the study of text types (Biber 1986 *et seq.*). With its 20 equally represented textual categories Cobuild's 17 million word corpus has still been criticized for being too 'journalistic' (Rundell and Stock 1992). As with LOB, Cobuild's corpus contains only random extracts of texts: there is not much room for a detailed analysis of the context of production or use of any of the texts even though extensive provision was given to allow for the bibliographic tracking of all citations.

Yet with the larger sizes of the second generation corpora, characterised as having used optical scanners to input text (Burnard 1992), 'representativeness' or an idea of what proportion of texts should constitute the 'norm' proved to be just as difficult. Analysing the differences between the main English language corpora (Brown, LOB and Cobuild) Ljung has found that within the first 1 000 words of each corpus 204 words were not shared. Ljung (1991:249) points to very important genre differences between the corpora, especially Cobuild, with its large number of high frequency abstract nouns to do with domains of behaviour, geometric shape and politics. Not much could be said about specific genres either: with their categories of text (around 12 000 words per genre in the original LOB corpus) the corpora were hardly more comprehensive than the number of words covered by manual analysis. Collins and Peters (1988) question the motivation behind the decision of many corpora (especially LOB) to give equal weight to texts such as 'belles

lettres, biographies and essays' as to 'the Press' or 'learned and scientific writings.'

The third generation of corpora including, in this country, Birmingham's (ex-Cobuild's) *Bank of English* and Oxford University and Longman's *British National Corpus* has been more quickly built up using access to electronic journalistic news files and other networks that have become available since the late 1980s. They have over 2 billion words of data each (Sinclair 1993a, Rundell and Stock 1992) with the original aim still to provide a representative sample. However, since the data they produce is still largely for lexicographic purposes the number of texts is still restricted to around one hundred million words to cut down the journalistic bias. Another notable corpus project, the Cambridge Language Survey is attempting to build up corpora and develop software for the analysis of seven major languages with particular emphasis on developing agreed codings for semantic and syntactic categories (tags) (Atkins, Clear and Ostler 1992). As lexicographic corpora grow, so do other types of corpora including general language 'core' corpora, dialectal corpora, grammatical corpora, spoken corpora, and specialised corpora (Svartvik 1992:12, Atkins et al 1992).

5.4 Corpus linguistics and differentiation of genres.

Because of the lexicographic race to provide a characterisation of the language in general, the study of specific text types has largely been of a stylistic nature involving word counts (Muller 1976). The lack of models or criteria for the selection of 'authentic' data lead Laurén and Nordman (1992:223) to report that there are "no models for the corpus selection of LSP research". There have been many studies of LSP texts on a very broad scale aiming at representative samples of 'technolects'. Laurén and Nordman argue that there have been few studies of specific genres because corpus linguistics, a field traditionally dominated by stylistics, makes no systematic distinction between register and genre, and typical work on stylistics, for example Oppenheim (1988), has largely concentrated on differences in word counts accounting for different authors' styles (Potter 1990:411). Most computational studies of style follow Enkvist, who has provided a definition of style that is tailor-made for the corpus linguistics community, being statistical in nature as well as incorporating the idea of instantial meaning:

> 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. (1964:28)

Among computational 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) studies discontinuous collocational frameworks in Spanish magazines and found that prose articles can be shown to be different to interviews in that the frameworks contain more textual information in the former and interpersonal, discoursal phrases in the latter. From the LOB corpus, Sampson and Haigh (1988) find that noun phrases, prepositional phrases, numbers of past participles and non-standard *as* clauses are more common in technical writing than in fiction but they argue against identifying "tell tale constructions" (1988:218). Gerbert (1970) has analysed 24 verb tenses in English technical writing, and finds, as do the genre analysts, that the present represents a limited set of meanings (scientific laws, processes and repeated actions, definitions, descriptions, observations and material properties). The perfect tense is used to indicate relevance to the research process.

5.5 The status of linguistic evidence.

Leech warns that despite its advantages in terms of parsimonious model building and strong probabilistic models, the danger of corpus linguistics is that it may not be well received in terms of *psychological plausibility* where corpus linguists cannot make any claim about the mental processes that are involved in the production of corpus data (1992:113). Renouf (1992) has also sketched the problems involved in taking metadiscoursal statements as part of a supposedly 'authentic' corpus. Whereas most texts in the Brown, LOB and Cobuild corpora involve situations where the context is partly reconstructible, the kinds of texts that make direct reference to the reader or the authors, or to other authors, opens up a wide degree of complexity. Renouf states that some codification of corpus data is necessary for signalling 'constructed spoken text' versus 'authentic spoken text'. There are however many divergent opinions as to the role of human involvement in the analysis.

The phraseological approach (Francis 1993) assumes that there should be as little human involvement as possible. All grammatical evidence should come from real examples analysed as automatically as possible as opposed to invented ones analysed intuitively or even introspectively. The principal research method of the Cobuild research group (Sinclair 1981 et seq., Francis 1987, Clear 1987, Krishnamuthy 1987, Renouf 1987) and researchers whom we might qualify as 'collocationalists' (Kjellmer 1984, Smadja 1993) has been the collation of words that cooccur in close cotext with other words.

Even linguists normally involved with the 'introspective' end of the linguistic spectrum, such as McCawley (1982) admit that when anomalies in linguistic data occur (the grammatical analysis of *Tom is having smoked pot* versus *Tom is having smoked pot* analogous to *Tom is having fried eggs*) it is largely a question of gestalt psychology, where one is "dealing with perceptual objects rather than graphic/real objects." (1982:79).

But many corpus linguists (Leech and Fligelstone 1992, Garside, Leech and Sampson 1987, Souter 1990) are involved in work that changes the format of the texts that they are working with, whether it is to transcribe prosodic markers from spoken texts, to full syntactic tagging (marking of word class and syntactic function) for the sake of parsing or empirical observation. Leech and Fligelstone (1992) consider that the counting of concordance items is at best "a trivial facility" and that the only significant data can come from annotated corpora. Similarly, 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)

Souter (1990) has observed the distribution of systemic-functional rules from a 100 000 word tagged spoken child language corpus. He found that 70% of the 8522 automatically identified componence rules are only used once in the corpus (componence rules are syntactic and functional phrase structures: such as Subject_NGP _ det head). 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). However, Briscoe (1990) has argued that although "all grammars leak slightly", there is no evidence for a group of 'deviant' grammatical constructs, arguing that one token of data on a particular rule does not rule out the applicability of the main componence rule.

Since most syntactic tagging systems fail to provide adequate tags for so-called 'discourse items', Svartvik (1993:24) has proposed a 170 tag system with labels such as *greeting, fluency device, hedge* and so on. 'Taggers' also favour lemmatization, the collective statistical analysis of alternative forms such as *be, is, are.* However, in his analysis of the 'velocity' or rate of change of frequency of new words in texts, Youmans (1991:766) finds that lemmatization does not significantly change the curves of type token ratios and is therefore unnecessary. In his analysis of the use of the word 'risk' Fillmore (1992) demonstrates that the word has a unique phraseology in the language in that 'running a risk' sees harm as a result of 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 alternatives. Chafe has taken 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)

Sinclair advocates a total reliance on data to the point where the corpus should not be grammatically tagged and forms should not be listed as lemmas (1991:7). One reason for this is that the findings of his colleagues at Birmingham (Fox 1987, Renouf and Sinclair 1991, Francis 1993) have tended to question categories that had been established by traditional grammars:

> 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)

Sinclair's argument is that traditional grammars have been concerned with grammatical competence and a notion of well-formedness, even for structures that appear to be unnatural because they are often too explicit. Yet authentic utterances appear to be unnatural because they have been taken away from their immediate cotext, making them seem either too cryptic or too implicit. Sinclair cites the following examples on a continuum from cryptical to explicit: *we searched, we searched all night, we searched all night for the missing climbers* (1984:206). Relations within the sentence that would lend themselves to a sense of the 'grammatically correct' are hardly relevant when one attempts to take account of the function of the sentence in context. The study of language has therefore to start off from a new approach to the notion of authentic data, so that while Sinclair criticises descriptions of language based on invented examples, his own corpus is based on texts that were not designed for lexicographic purposes. The idea that there are patterns that may not be in

accordance with previous theory leads him to a superficial approach to corpus work: the analysis should accept and reflect the evidence at the cost of theoretical elegance as Halliday suggests: "system and instance are not two distinct phenomena [...] " (1992:66).

Similarly, Church and Mercer (1993:4) state that parsers, which impose tagged structure on text, are useful for 'understanding who did what to whom,' but are less useful for predicting likely usages in authentic language. The other disadvantage of parsers and tagging systems is that they have, according to Church and Mercer, little success in word class or word sense disambiguation (1993:9).

On the basis of these arguments, Francis (1993) has also argued that grammar should be based on lexical structures as opposed to syntactic ones which see lexis as simply data to be fitted into slots. Halliday (1966:50) had raised this issue in his discussion of the adjective-noun pairs *strong tea* and *powerful computer*. Clearly they are not interchangeable yet there is little in the systems of syntax or semantics that can account for their difference in distribution. Halliday's concept of a lexicogrammar and the relations between the grammatical and lexical systems are discussed by Willis (1993) who suggests that the phraseological approach leads to a fragmentation of grammar and a move away from a grammar of word classes to a description of lexical behaviour. The highest levels of grammar have been the most extensively explored, namely the 'grammar of structure' indicating rank scale of subject, predicate, object, and complement, and the 'grammar of necessary choice' (Willis 1993:85) including the kind of large scale bipolar choices such as mood, tense and aspect as set out by Halliday. Willis (1993:87) points out corpus evidence that some verbs such as *be, think* carry different meaning with continuous aspect, or have a very restricted number of possible objects (*entertain this idea, effect an entry*). These observations demonstrate how a grammar of structure and choice is limited to seemingly *ad hoc* lexical statements. At a lower and less well known level is 'the grammar of class', the behaviour of items at the level of lexis, where the most delicate information available tends to be 'uncountable noun' and 'verb with no continuous aspect'. Willis also claims that corpus evidence shows how classes merge into one another and how some high frequency words have unique syntagmatic environments, while subsets of word classes have very different properties to the traditional class as a whole. For example, many nouns modify the semantics of common 'delexicalised' verbs (*give a smile, take a chance*) or are involved with clause structures with *that* (*belief, argument*) infinitives (*decision to, claim to*) and complex nominals with *of* (*behaviour of, arrival of*). Finally Willis claims that there is an even less well understood category of lexical behaviour that involves the rhetorical

construction of discourse, using structures like *the main/important/other thing/ question / problem / difficulty is that...* plus some statement of a problem and where a following solution is signalled and needs no further modification or evaluation (1993:88). Willis terms this the 'grammar of probability and collocation' and 'structures of discourse' (1993:89).

5.6 Phraseology and the Idiom Principle.

From a terminological perspective that sees reconceptualisation as a dynamic force behind the scientific innovation of paradigms, and from lexicography where the phraseological principle has been established as a satisfactory way of describing the lexicon, the concept of collocation arises to form in this thesis a central area of interest for the study of scientific text. As far as descriptive techniques of language go, the post-Firthian concepts of collocation and phraseology (and indeed colligation) have only recently emerged as acceptable alternatives to syntax.

Sinclair's view of collocation and its role in the description of English has been influential in the world of lexicography where Sinclair's work has seen the more tangible areas of application in his editorship of the corpus-based Collins Cobuild dictionary (Sinclair 1987a). Sinclair's idiom principle shares its roots with Firth's original conception of language as a system and Halliday's description of lexis as delicate grammar:

> 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, its own distinct realization [...] At the other end are much more specific, loose, more shifting sets of features, realised 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 starting point is a rejection of the 'open choice' principle. According to this model of language, any lexical item may be fitted into a slot that fulfils a certain syntactic function within a syntagmatic structure. This grammar consists of a series of rules that include rules for transformation and internal variation derived from a basic deep cognitive structure and based on logical principles of constituency (Chomsky 1967). Sinclair's main argument against the open choice principle is that it is unlikely that normal text would be produced if we were simply using these criteria. If the open-choice principle really applied, then

colourless green ideas would indeed sleep furiously- there would be little evidence of recurring patterns of language.

The idiom principle, on the other hand, supplements the criterion of grammaticalness (Sinclair 1987c:320) by a criterion of naturalness. It's not enough to describe an utterance as grammatical: it has to be perceived as natural by a native speaker. Idiomaticity relies essentially on the evidence that words tend to co-occur, thereby reducing a vast amount of possible combinatory possibilities, the combinatorial explosion that would be allowed in an open-choice system. Instead of open-choice, language is seen as a phraseological system, where clusters of more than one word reflect a single lexical choice. One set of conditions that may reduce or even pre-empt lexical choice even further may be attributed to register (1987a:320). But Sinclair claims that even register does not restrict the open-choice principle enough to account for the high degree of idiomaticity that he expects. Since language is an autonomous system, the way the world is organised through language is not just a one-to-one translation but a system with its own in-built criteria:

> 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 analyzable 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. (1987c: 320)

Sinclair argues that the most tangible evidence of larger-than-word items is the prevalence of proverbial expressions, clichés and technical terms in all types of discourse. In particular compound items have lost their status as separable words (*because, of course, maybe, another* and so on). To this we can add what Nattinger and Decarrico (1992:24) and Willis (1993:88) refer to as holophrastic phrases: prefabricated chunks of language that have a very stereotyped usage in discourse, as in child discourse: *wanna + VP, allgone, what's that?* /hw_dæ/ and so on.

Similarly, high frequency content words (such as verbs as *get, set, take*) also have less clear semantic profiles, where the most useful definitions for these words cannot avoid examples of usage, that is their typical cotext. Since these words constitute the majority of any text, then a normal text can be characterised as being largely "delexicalized" (1987c:323).

5.7 Phraseology and discourse.

In this section the idiom principle is further explored in the context of a phraseological approach that has been advocated by Sinclair's colleagues and which now constitutes a broad critique of opposing models. A phraseological model itself has not emerged, although researchers such as Nattinger and Decarrico, Willis and Moon have attempted to relate collocational principles to discourse functions. The Cobuild group especially expands the idea of an individual 'fact' about language (perhaps the kernel idea of phraseology) to a number of collocational phenomena that might be termed discourse items. Francis's argumentation stems from a variety of seemingly unrelated issues arising from the pre-emptive properties of semi-idiomatic phrases as in *put a brave face on it*, semi-prepackaged idioms with clear communicative goals (*not the foggiest/faintest idea*) and prefacing items (such as *be/is a case of*) where a current discourse topic is compared to one familiar to the reader (1993:143-6). The use of such phraseological items hardly seems to fit into a structural paradigm of grammar, nor has it been fully realised in systemic grammars. Francis also uses a number of corpus based examples to demonstrate how high frequency pronouns, conjunctions, adjectives and nouns have a unique phraseology (1993:140), a different line from that taken by the *BBI* dictionary (Benson et al. 1981) which eliminates common words (such as *big, cause* and *make*) that can, according to Thoiron and Béjoint associate with "almost any words in the language" (1992:7).

Francis cites the example of adjectives and adjectival clauses in *-ing, that* and *to* in complement position that, according to Quirk et al. (1985) may undergo extraposition with *it* as object followed by an adjective or noun group as in *they often find it difficult to explain why*, with a possible un-extraposed reformulation being *they often find explaining why difficult*. Whereas Quirk et al. do not identify lexical patterns associated with this structure, Francis finds in the Cobuild corpus two main lexical structures that occupy 98% of all occurrences of the extraposed *it* structure: *find* and *make,* followed by a very restricted set of adjectives (concerning the concept of ease and probability) or structures such as *make it clear/likely that.*

Francis also finds clear patterns (1993:46) for the word *possible* where it mostly occurs with superlatives, in the frame *as X as possible* and after *whether/if...* Similarly, appositive *that* after noun groups (as in *the idea that, the advantage that, the chance that*) has a series of structures that can be enumerated according to lexico-semantic range such as illocutionary processes (*allegation that, contention that*) and thought processes followed by their results

(*analysis that, realisation that*) (1993:149). Appositive *that* can be analysed at a higher semantic level, and Francis identifies the left-collocate as identity encapsulating what is to be introduced and the right-collocate as some kind of explanation (hence: *his forecast that there would be snow.*) Since the communicative function, Francis claims, is to set out an evaluation based on coping with some situation which is detailed after the verb, the structure cannot simply be described as a syntactic opposition between extraposed *it* clause and clause as object. Francis argues that this phraseological knowledge is just as essential as the syntactic definition given by traditional grammarians, and may well be more accurate. On a more syntactic level, Francis lists "phraseological constraints" for *reason* (1993:152) that alternate between phrases with *that* and *for* and a fixed phrase *for the simple reason that* as an emphasising precursor.

Francis predicts that high frequency grammatical items like *that* and *it* or lexical ones like *possible* and *reason*, as Hoey has demonstrated (1993), will not have the same phraseological constraints as typical word classes (conjunctions, nouns and so on) (1993:147). As we see in the Conclusion (Chapter 12) Francis claims that if an item becomes more frequent, its occurrences become more idiosyncratic, since its collocational properties will be spread across a variety of functions. The traditional word classes have less importance, therefore restricting the value of grammars that use classes as a primary level of structure. This communicative approach to phraseology provides a powerful account of a single conceptual choice in motivating linguistic structure. Sinclair states that it is this strategy of communication that entails initial lexico-grammatical ramifications and that "grammar is part of the management of the text rather than the focus of the meaning-creation". (1991:8). Since the collocationalists see structure and communicative use as inseparable, the idea of a single choice that mediates its own syntax and structure automatically on the basis of a few targeted lexical items is a powerful one. Whereas in the traditional generative paradigm structure is seen as a medium for meaning, allowing meaning to be encoded and decoded, the collocationalist sees the single lexical choice as the basis for what one might term a cascade of obligatory phraseology that follows on to further choices in the system. Once the conceptual choice is made, the communicative constraints on phraseology are autonomous (Sinclair 1991:7).

5.8 The Structure of phraseological collocations.

Bengt and Altenberg (1990) have identified two main currents in the study of the structure of collocations: firstly the broad syntagmatic sense of "word sequences in text" (idioms, compound, complex words) that forms the main approach of lexicography and traditional linguistics. Secondly there is a narrower sense of "lexical items in the language" (1990:3) which cuts across the traditional word boundaries and clearcut sense of grammar and lexis. Some of these ideas have been partially accounted for and precipitated by the Cobuild and Longmans dictionaries. As well as the syntagmatic views of "mutually occurring lexical items" there is also the stronger hypothesis of "mutually selective lexical items" proposed by Bengt and Altenberg (1990:3) and Cruse (1986). Finally there is the view that collocation depends on both structural and intuitive criteria, where collocation is defined in syntactic terms (N+V collocations) or in terms of 'fixedness' (Cowie (1981), Benson (1988), Howarth (1991), Kjellmer (1984) and Leech (1992)).

There has also been evidence from corpus linguistics to support the claims of the idiom principle and the lexicalist position. Kennedy (1990) has reported that 63% of the use of *at* is limited to 150 collocations, with *at least* being the most frequent. Krishnamurthy (1987:70) reports from the Cobuild corpus that many common items have very restricted collocations, such as 70% cooccurrence of *refer* with *to*, while 100% use of *encrusted* as an adjective or past participle and *backsliding* as a noun rather than a verb. Altenberg (1991) has argued for types of collocation not normally associated with idioms and word compounds. One such is the 'amplifier collocation' type of intensifier such as *absolutely* which occurs with superlative adjectives, and *perfectly* which collocates with negatives (*no*, *not*) or very positive statements.

Not all collocationalists share the phraseological 'long range' view. An opposing view may be formulated on the grounds that the corpus can never throw up all the new combinations that are possible given certain syntactic principles. Kjellmer (1984:163) argues that there are either randomly thrown up recurrent word combinations (*although he, hall to*) and unusual grammatically restricted sequences (*green ideas, yesterday's evening*) and that at the intersection of the two are valid phraseological units (*last night, try to*). From the tagged Brown corpus, Kjellmer (1990) finds statistical 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 noted also by the Cobuild group (Krishnamurthy 1987).

While not contradicting the phraseological perspective, Moon (1987) has suggested that an overemphasis on context, especially with high frequency words, has led to an overabundance of meaning distinctions where, in lexicography at least, one runs the risk of "losing the semantic integrity of the word." (1987: 102). On the other hand, there often appears to be too little context for words that express discourse or clause functions (*and, but, however*) or collocations that appear to require quite a large cotext such as (*so ...as*) as Kaye (1990:151) notes. Moon argues that corpus evidence for items which might have millions of citations like *the, a, this, that* and *such* overlooks their basic cohesive features. She suggests a more dynamic view of collocation in discourse, especially of idioms, that attempts to classify collocation not just on syntactic grounds, as Kjellmer, but also on the basis of their intended use in discourse "the discoursal function of a fixed expression may be defined as the contribution it makes to the information structure of a text." (Moon 1992:495). She distinguishes between *pure idioms*, which are institutionalised metaphorical phrases of varying semantic transparency: (*fly on the wall*), *anomalous collocations* which are grammatically idiosyncratic or semantically opaque (*might as well, for once*) and *restricted collocations* which are phraseologically regular but otherwise less lexically marked (*take steps, from side to side*). Anomalous and restricted collocations take up 85% of Moon's corpus of 18 million words. And their syntax in 75% of cases is of the form predicator (i.e. verb) + object/adjunct or some form of adverbial or adjectival group or complement as a whole. In terms of discourse function, Moon states that 50% of the fixed expressions she found were informational (*for sale, rub shoulders with, in the running*) while 35% had more evaluative overtones (*kid's stuff, at a snail's pace, near the knuckle*) and 10% conveyed some truth value (*I kid you not, you know what I mean, to all intents and purposes*) (1992:496-497). The other two categories she uses (reflecting context and signalling the discourse structure) account for 5% of her samples.

Nattinger and Decarrico (1992) claim that discourse signalling is an important characteristic of a group of phraseological forms they term lexical phrases and that they formally distinguish these from idioms, clichés and collocations. They set out a classification of phraseological phenomena that sees the communicative lexical phrase as a more useful unit for pedagogical purposes and extend their view of collocation beyond statistically

predictable collocation. Their view approaches Pavel's terminological collocation in that they link the lexicon to linguistic knowledge of the world rather than to the behaviour of words (1992:22), a significantly different perspective to the Birmingham school. 'Syntactic strings' (such as NP +Aux +NP) are created grammatically by syntactic competence. Collocations are seen as significantly co-occurring lexical items that "have not been assigned particular pragmatic functions by pragmatic competence" (1992:36) similar to the types of cooccurrence noted by Halliday (1991a, 1992) and others. Lexical phrases are collocations that do have pragmatic functions and can be split into two groups (1992:38-42): those that do not allow paradigmatic or syntagmatic substitution with 2 subtypes: polywords: *for the most part, as it were* and institutionalised phrases *how are you? what, me worry?*). Secondly there are frames with both fixed and free features: short term 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]*). Nattinger and Decarrico claim that the lexical phrase is an important linguistic unit in language acquisition and that it requires more attention in the teaching and learning of native-like discourse:

> Lexical phrases are parts of language that often have clearly defined roles in guiding the overall 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 has preceded. (1992:60)

The communicative functions that Nattinger and Decarrico propose for lexical phrases include social interaction, where conversational maintenance and purpose are signalled, phatic communion and social necessities and discourse devices which in turn include logical connectors (*as a result*), temporal connectors (*to begin with*), exemplifiers (*in other words*), summarisers (*that's all there is to it*) and various types of evaluative expressions such as integration (*most researchers agree that*) and detachment (*it has been noted that*) (1992:61-85).

On the basis of a 10 million word corpus (from Associated Press) Smadja (1993) has established four principles of phraseological collocations:

-Principle 1- Collocations are arbitrary (1993:146).
Collocations are combined on a local basis that may not have any semantic or syntactic explanation. This can clearly 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 can have a very local 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. (Church and Hanks 1989)

-Principle 4- Collocations are cohesive lexical clusters (1993:147)
Collocations are internally consistent and often have internal elements that are predictive of others. Although Smadja claims that this is unlike Halliday's textual definition of cohesion, there is clearly a sense of unity and 'texture' that Halliday and Hasan (1976) refer to within collocations such as *heavy trading,* or *agree to.*

However, Smadja's computerised system for determining collocations eliminates grammatical items and collocations which do not have an identifiable syntactic structure (1993:406). This leaves three types of collocation, which corresponds exactly with Burnard's classification (1992:14):

Type 1 collocation- 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 the basic meaning (Smadja 1993:399). These correspond to Cowie's (1981) and Benson's (1988) restricted collocations.

Type 2 collocation- Rigid noun phrases.
These are "important concepts in a domain." (Smadja 1993:148) such as *stock market* and *Dow Jones* and have been established by Choueka et al (1983) in their study of the New York Times corpus and by Burnard (1992:15) who terms them 'text-oriented' cooccurences.

Type 3 collocation- Phrasal templates.

These are collocations which include very free elements within a restricted structure (such as *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's statistical approach and Nattinger and Decarrico's intuitive approaches exemplify two models for identifying collocation. The first is highly satisfactory in empirical terms, yet reveals a very poor set of data although Smadja is limited to word classes by his use of a tagged system of elimination of unwanted collocation. On the other hand, the intuitive method appears to have all the prejudicial disadvantages of other intuitive methods of analysis (for example semantic) and yet reveals a complex system of long-range discoursal signalling that would not surface with ease from a statistical view of collocation. The most established theories of collocation belong to the field of lexicography, from which Cobuild is but one of many projects exploring and applying phraseology. So while lexicography has little relation to the ultimate aim of this thesis to characterise scientific text, the experience of lexicography may aid the development of a new methodology, or at least to understand the strengths and weaknesses of existing methodologies.

5.9 The structure of Collocations in LSP

Benson (1986) offers an alternative vision of collocation and establishes some fundamental differences between LGP collocations and LSP collocations. He notes that in the LGP, adjectives tend to be used in such phrases as 'strong tea, best regards' and 'formidable challenge' and tend to introduce infinitive clauses with an adjective 'it's *good* to see you', whereas in the LSP adjectives tend to be past participles such as '*restricted* enzyme, *attenuated* strain' and tend to be used to introduce clauses with 'that': 'it seems *possible* that...'. Benson (1986) also notes that in LGP compound nouns where the elements become more specific such as 'cabinet *reshuffle*, drug *pusher*' and 'brake *booster*', the attributive nature of the second element can be reinforced by deconstructing with 'of': 'a *reshuffle* of the cabinet, a *pusher* of drugs, a *booster* of brakes'. However with specific-generic terms such as '*measles* vaccine, *jet* engine, *house* arrest' such deconstruction is unproductive. Since the second set of forms are more typical of the LSP, Benson argues, LSP nominal groups must have a generic-specific internal structure that distinguishes them from their LGP counterparts. The lack of reformulative potential of a multiword term therefore allows a distinction between fixed LSP terms and LGP phrases. Salager-Meyer

(1990a:354) has also reported that 70% of head nouns in medical terminology tend to be metaphorical collocations involving structures (nerve *roots*, abdominal *walls*) and while the rest involve processes, functions and relations (migratory *pain*, *vehicles* of infection).

Thomas (1993) provides further evidence of specific patterns of collocations in LSP multi-word terminology when she describes the types of collocation that occur in a computer based terminological term bank. Thomas 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 goes from left to right. Thomas notes that collocational variability, where the node is highly predictive of the left or right collocate, affects the lexicographer's choice of base word and Sinclair refers to this phenomenon as a statistical problem of 'up or down direction of collocability' (1987c:330). Sager et al. (1980:231) suggest that potential collocation in the general language is freer than in the special language.

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. also 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...*) introduces additional thematic elements at the beginning of the sentence. In our analysis of posture, below, we find this is a typical, and rather idiosyncratic device for prospection (prediction) in the research article genre. Gerbert (1970) has also characterised sub-technical verbs in technical prose (items like *effect, assume, result in, increase, measure*) as "semantically neutral" (1970:38). Thomas finds in her analysis no examples of phrasal verbs in terminology (such as 'put up') (1993:54), although this presumably depends on her definition of 'phrasal verbs' (for example, *to boot up, to log in* and *log out*). She also surprisingly states that "prepositions ... do not appear to play a role in LSP phraseology." (1993:64), although this may have some connection with the size or nature of her corpus. Benson's specific-generic

pattern also applies to Thomas's view of *noun* (functioning as adjective) + *noun* collocations for example, in the LGP phrase *drug pusher* the last item is specific/determining while in the LSP *polypeptide chains* the last item is generic/determined and is not naturally deconstructed to *chains of polypeptides* (1993:53).

PART II: SURVEY DATA COLLECTION

6.0 CHAPTER SIX: Research Synthesis.

In this chapter, the ideas introduced in the literature review (Part I) are synthesised. The primary and secondary research questions from Chapter 1 are addressed here by specific research hypotheses. The primary hypothesis is addressed in Part II in terms of a survey. The two secondary hypotheses are tested in Part III: Data analysis. Their results are integrated with a discussion of the primary hypothesis in Part IV: Conclusions.

6.1 Research Synthesis.

Let us recapitulate the main thrust of the literature review. This thesis centres on how scientific innovation is enacted by cancer research articles and on how language functions in very specialised scientific research activities. On the basis of this we formulated three research questions:

General question: What role do research articles have in scientific activity?

Specific question 1: What role does the research article have in the creation of new science?

Specific question 2: How does language function in extremely specialised contexts?

We answer the first question in Chapter 7 below. Here we concentrate on the two specific research questions.

Discourse analysis has suggested that text plays a key role in the social construction of science. In particular, the field of ESP has demonstrated the rhetorical nature of science writing, and we have seen attempts to describe the rhetorical structure of scientific text (Swales 1990) and the rhetoric nature of scientific innovation (Myers 1990). Specifically, Halliday and Martin (1993) have proposed the textual progression of grammatical metaphor as a mechanism for change in scientific text. More generally, Sinclair has proposed a dynamic model of discourse signalling which may shed light on the processes of textual argumentation. Such a model depends on the use of discourse items which have been proposed by Tadros (1985) and Francis (1985). We consider in this thesis that grammatical metaphor and discourse signalling constitute different aspects of a single linguistic mechanism: reformulation. Our first hypothesis is that reformulation is a key process in the construction of scientific claims.

At the same time, computer-corpus research has provided a methodology for arriving at a high degree of descriptive accuracy for large amounts of authentic data (Sinclair 1991). This work has already been related to the study of scientific innovation; most notably by Pavel (1992) who has, as we have seen, emphasised collocation as a key process in the way a scientific text creates neologisms in contrast with existing preferred expressions. Similarly, Moon (1993) has proposed that idioms have their own specific discourse functions in running text. We consider collocations to act in conjunction with rhetorical functions and this constitutes the linguistic principle of 'phraseology'. The second central claim of this thesis is that language in scientific text not only displays highly conventionalised phraseology but functions by reformulating it.

We have seen that the 'reformulation' approach has not been applied to large numbers of texts. Conversely, the 'phraseology' approach has not been applied to specific corpora. We have argued, then, that it would be profitable to exploit both. To summarise, the two specific hypotheses explore the following phenomena in cancer research articles:

-grammatical metaphor
-discourse signalling
-collocation

6.2 Research Hypotheses.

6.21 The Adaptive Science Hypothesis.

General Research Question: *What role do research articles have in scientific activity?*

While scientists may not perceive language to be a matter of priority in their work, they are conscious of the need to sucessfully present and disseminate their ideas, particularly in competitive publishing. Just as much as scientific text is written to allow readers to access it as a reference work or index, it is written to persuade. The constructivist paradigm, which this thesis adopts, claims that science only resides in language as a socio-rhetorical construct. But the paradigm does not deal with how new terminology is textually determined or how innovative ideas are constructed throughout texts. Science is defined as a linguistic concept, but there are no indications about how language is used to bring about change in science. We are not concerned with how new science is successful or enacted; that is a matter for the rhetorician. Instead, this thesis is essentially concerned with the

mechanisms of language that are at stake in reformulating ideas in research articles. The following hypothesis provides a theoretical backdrop to the specific research hypotheses, below:

The adaptive science hypothesis: Science writing (in particular the research article) is the only means by which scientific change can be enacted. The linguistic resources used by scientists to enact change are specific to the discourse community, although they have been adapted from the wider language system.

Testing the adaptive science hypothesis: The hypothesis seeks to supplement current debate on 'science as a social construct' by testing the idea that scientific knowledge actually only resides in linguistic form. We can partially demonstrate this by conducting an ethnographic survey, analysing in depth the processes of use and production of scientific texts in a particular discourse community. But we require essential evidence from the linguistic analysis of the two specific hypotheses in this thesis. It is for this reason that the results of the survey are discussed summarily in Chapter 7, and further discussion of the adaptive science hypothesis is put off until Part IV (Conclusions) where the two specific hypotheses can be integrated in the light of the survey and corpus findings.

6.22 The Reformulation Hypothesis.

Specific Research Question 1: *What role does the research article have in the creation of new science?*

We claim that the 'indexical' function in cancer research articles extends to the language of cancer research via a general property of language: to encapsulate and predict past and future discourse. The aim of the reformulation hypothesis is to reveal the key processes involved in research writing. Rhetorical structure is one way that has been explored, and we discussed some aspects relating to ESP and Genre analysis in Part I. In this thesis we concentrate on two possible mechanisms that operate on a more local basis than rhetorical structure: the reformulation of claims and the creation of new expressions. Reformulation may function by using expressions from the existing discourse community and reformulating them as the text progresses as grammatical metaphor (Halliday 1985, Halliday and Martin 1993, Derewianka forthcoming). These grammatical metaphors may be combined in ways that Pavel (1987) has suggested, allowing new collocations to emerge. To identify the role of texts in cancer research the following hypothesis is tested:

The reformulation hypothesis: New ideas in cancer research are expressed by the interaction of two textual processes: 1) grammatical metaphor and 2) posture (with particular reference to discourse signalling).

Testing hypothesis 2: The hypothesis is that while rhetoric may account for the effectiveness of a new argument, the linguistic form the argument takes involves gradual rewording from an established formulation to a new one. This hypothesis can be tested by the analysis of reformulation in a sample of texts and the comparison of reformulation and grammatical metaphor in a smaller sample. The hypothesis is falsifiable if it can be proved that there is no recurrent pattern of reformulation in posture or grammatical metaphor or that these processes do not correspond to the construction of scientific claims.

6.23 The Phraseology Hypothesis.

Specific Research Question 2: *How does language function in extremely specialised contexts?*

The assumption of this thesis is that phraseology is a useful model for the interaction between the linguistic system and the rhetorical processes involved in writing and reading. Since computational techniques have been developed to explore phraseology in large corpora, and since this has not been fully exploited in genre analysis, the following hypothesis is considered to be the central research emphasis of this thesis:

The phraseology hypothesis Collocational patterns correspond to rhetorical functions, and collocational patterns are consistent within rhetorical sections of cancer research articles.

Testing Hypothesis 3: The hypothesis is that collocation in the whole language system of the genre varies according to rhetorical section. In the case of this thesis, each title, abstract, methods, results and discussion section is considered to share constant global rhetorical characteristics with rhetorical sections of the same name. The hypothesis can be tested by the following condition: If collocational patterns are related to rhetorical purposes, a significant difference should be noticeable between collocational patterns in titles, abstracts and the other major rhetorical sections of research articles. Phraseological analysis should reveal the most typical rhetorical functions of each rhetorical section and link each key rhetorical expression with one phraseological pattern.

6.3 Research Methods.

Hypothesis One (the adaptive science hypothesis) is tested by a survey. The reasons for choosing informants from the Department of Pharmaceutical Sciences are described in Chapter 7. The reformulation hypothesis requires a number of texts that can be elicited by the expert informants from the survey. Since one purpose of the survey is to elicit information about the way scientists use research articles to access or to produce new science, these findings can be used to support observations for the reformulation hypothesis.

The phraseology hypothesis requires a representative corpus of texts. The Birmingham Cobuild and Longmans' dictionaries have found that large amounts of data are required to describe common language principles. However, since the phraseology hypothesis is directed at an extremely specialised genre, even a small corpus would be as relatively representative as the early corpora used by lexicographers. Selectional criteria for the corpus are necessarily as complex as those for the larger general language corpora. Since a survey is carried out for the adaptive science hypothesis, the corpus needs to reflect the research interests, goals and activities of the experts who can then answer queries that arise from further analysis.

Essentially, the three hypotheses attempt to link purely qualitative information from a survey of expert informants' opinions with quantitative linguistic data. If genre analysis is to make any use of corpus linguistics, this association needs to be justified. Further, the methods of observation and ways of eliciting opinion need to be elaborated. The naturalistic approach, set out in the section below, provides a key to how to conduct such a survey, maintaining that the values of both informants and linguists need to be recognised in the analysis.

6.4 Evaluation of the available methods: Guba and Lincoln's Naturalistic Approach.

As mentioned in the introduction to this thesis, Guba and Lincoln (1992) argue that a rationalist paradigm is unsuitable for conducting qualitative research in the humanities. They maintain that observations of social phenomena are necessarily skewed because the observer projects his or her own social values onto the activity in question while the

'subjects' or expert informants similarly react to the act of observation. According to Guba and Lincoln, the enquirer should act as a *smart instrument* (1992:240); able to adjust the search strategy and judge informants' responses in the light of the social interaction. Also, social phenomena and truth statements should not be generalised into monolithic rules, but can only be situated in an *ideographic body of knowledge* (1992:238) to be regarded as temporary and dependent on context. Thus a concept need not be defined in terms of pure semantic components, but in terms of its values to researchers, and its implications of central and peripheral concepts and methodologies. Guba and Lincoln suggest that difficulty in obtaining such a body of knowledge is an interesting result in itself and the ways in which an ideographic map is arrived at are as empirically revealing as the phenomena under scrutiny.

However, any research must be able to go beyond the specific case in question in order to provide replicable generalised statements. In order to obtain 'internal validity' for any study, Guba and Lincoln suggest that the observer be as close to the social action as possible (1992:247). Thus to ensure that the results of any working hypotheses are transferable to other studies, they propose *thick description* as a means of assuring that the researcher is familiar with the subject matter and local issues his or her subjects are involved with. *Thick description* is detailed description that takes account of all possible factors and attempts to correlate rather than prioritise or classify them. For us this entails analysing and respecting the nature of scientific knowledge as well as the activity of scientific research. And to ensure the central empirical principles of 'dependability' and 'confirmability' Guba and Lincoln propose *triangulation*: the use of different methodologies to be compared with the enquirer's value-judgements to test against, and to rule out inconsistency (1992:246-8).

This context-dependent approach is entirely in accordance with the philosophy of the Firthian school of linguists, as discussed in the literature review, and Guba and Lincoln's view of a social system is similar to Firth's view of meaning construed within a system. Guba and Lincoln's naturalistic paradigm not only sees social systems in terms of a conceptual space of totally interdependent variables but also as a way of accounting how these concepts acquire meaningful value, as they say:

> The phenomena we deal with cannot be touched, seen, tasted, smelled or heard. That is not to say that tangible objects, events and processes do not enter into human behavior, for example, to shape it. However, it is not these tangibles that we care about, but the meaning and interpretation people ascribe or make of them, for it is these constructions that mediate their behavior. (Guba and Lincoln 1992:239).

Other researchers looking at the related phenomenon of abstract indexing have adopted a similar stance. Sharp (1989) sees her analysis as holistic, incorporating the techniques of professional indexers' experience into an expert system, while Gibson (1992) adopts the multivariate approach using many systems of description of experts' subjective data to provide triangulation. Selinker, Tarone and Hanzeli (1981) have established similar precedents in their use of expert informants as a support for their linguistic data.

The recommendations of the naturalistic approach to the study of social phenomena (*the observer as smart instrument, the ideographic body of knowledge, thick description, triangulation*) clearly have direct consequences for the survey to be carried out in this thesis. Any description should take account of the everyday research practices of the authors and consumers of abstracts and research articles.

Linguistic features of abstracts, for example, must be analysed as part of the local set of goals of the writers as well as of a wider system of scientific discourse. But the naturalistic approach also requires a pre-analytical justification and description of the subject matter in terms of its value not only to society and its immediate users, and where they see they have a part to play, but also to the linguist. This is addressed in the first part of the methodology in Chapter 7.

7.0 CHAPTER SEVEN: Research Context.

In the introduction to Part I of this thesis, five initial reasons were given for studying the research activities of the Department of Pharmaceutical Sciences at Aston University. Here they are expanded with an introduction to the department, and with details of how the specific research activities of the department are to be exploited in the survey and corpus. The analysis of the survey are discussed in detail here, especially those aspects of the research context which have particular relevance to the construction and description of the corpus. More specific findings are summarised in the survey details (Section 7.7) and are integrated into the discussion of the corpus in Chapter 8 and in the Analysis section (Chapter 11).

[From this point, in order to differentiate their opinions, researchers are referred to by their italicised initials (as listed in Appendix A). 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 article titles and bibliographic data which are also listed in Appendix A.]

7.1 Accessibility to the Department of Pharmaceutical Sciences.

The fact that researchers in Pharmaceutical sciences were easily accessible and interested in the role of language in their work was a major factor in gaining permission to interview them. The researchers were also willing to demonstrate their routine in the laboratories, and gave free access to written research and publicity material, including departmental listings and press cuttings. They were also happy to see that their activities aroused interest in other parts of the university.

None of the researchers had time to undertake more than one formal interview (lasting usually one hour); and so it was decided to survey as many researchers as possible in order to get a broad 'snapshot' view as opposed to the very close longitudinal study of the type undertaken by Myers (1990). Fifteen people took part over a period of six months. Even though the number of people interviewed includes only a third of the current academic staff in the department, the research activities of the whole department can be considered to be reasonably covered by this survey, since many of the researchers represent other members of small teams (three to four people publishing series of papers together).

7.2. Research Profile of the Pharmaceutical Sciences Department.

The department is a grade 4 research establishment (5 being the maximum at the time), with a high output of research and with a number of high profile breakthroughs. 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 the targeting of disease by the development of highly specialised synthetic compounds (organically functional substances that are artificially produced). This conceptual difference is represented in an institutional division between three departmental sections (pharmaceutical - cancer research - applied pharmaceutical) and five research groups. In 1992 the size of these groups (not including postdoctoral workers and technicians) was as follows:

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

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

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

When the researchers were asked to donate research articles for the corpus, the aim was to establish a distribution of papers that would represent not just pharmaceutical sciences, or cancer research, but the area of interaction between the two. The heterogenous nature of the specialisms, and the differences between prestigious, high impact journals are issues taken up in Chapter 8. It is only necessary to point out here that the high degree of overlap between the specialisms, especially within section II, means that the different research specialisms represented in the survey can be regarded as points on a continuum, rather than in opposition.

As well as a constant record of publication and successful applications for grant proposals, the department had a number of 'breakthroughs' during the period of the survey which on one occasion reached the attention of most of the local press as well as national press. Since the media pay special attention to cancer research, a large number of press cuttings were available on the basis of *MT's* work for the charity *Cancer Research Campaign* (CRC). He stated that it was a matter of CRC policy to keep a constant stream of 'breakthroughs' in the press, thus enabling different research establishments to gain funding. This is also evinced

by the frequency of 'cure for cancer' stories that are reported on television and in the national press. The fact that the research goals of the informants were not static also meant that researchers were frustrated by lack of recognition: the pharmacist *SF* had been obliged to switch his research to DNA molecules from his more original work on a specific inhibitor because of departmental policy.

Such matters of policy and presentation presumably constitute an area of tension in the department, and are an important factor in the description of the goals of the different research groups. In an environment where pharmacists are competing for research funding from cancer research organisations at the same time as cancer researchers proper, perceived relevance of specialism must have a consequential effect on a researcher's place in the hierarchy of his or her field.

7.3 Cancer Research within the Pharmaceutical Sciences Department.

While the main fields of expertise in the Pharmaceutical Sciences department are concerned with medicinal applications of chemistry to a number of major diseases (including rheumatism, AIDS, tuberculosis), the largest research group in the department is the Cancer Research Group, which maintains its own identity, including a visitors' suite and separate logo. Cancer research has a high public profile in terms of the nature of the disease, charity fund raising and publicity. The field is served by a large number of journals that are entirely devoted to it (as opposed to journals dedicated to other diseases, or fields such as microbiology or pharmacy) and the disease is regularly featured in a large selection of the most prestigious journals.

While the Science Citation Index lists over 8 000 journals of all the empirical and physical sciences, medicinal applications of biochemistry account for two thirds of the first 100 on the list. Of the first 600 journals on the SCI list (SCI 1993), 18 (3%) have cancer or oncology in their title while only 2 cite AIDS, and other diseases have only one journal specific title each (arthritis and rheumatism, heart disease, leprosy, schizophrenia, *inter alia*).

Thus medical science appears to dwarf other areas of scientific research. And given the large number of other medical fields, cancer research can be seen to be one of medicine's most prominent objectives. However, this may simply reflect the different culture of cancer research compared to other pressing medical issues: it has already been noted that a constant

presence is encouraged and maintained by frequent publication and a high profile media campaign. Attempts to claim centrality in such a large research programme are clearly extremely difficult. It would therefore be particularly fruitful to understand the dynamics of the discourse of cancer research, to study the ways individuals and groups of researchers attempt to enter this discourse from the apparent obscurity of their particular specialisms and to determine how they gain attention and claim relevance.

7.4 The Discourse of Cancer Research.

Cancer research itself covers a broad sweep of specialisms (drug synthesis, genetics, patient care) that are integrated into the global aims of the researchers within the institution. The various research activities (chemotherapy, metabolism studies, causal nutrition studies) contribute to solutions leading to the ultimate goal: the cure for cancer. Some researchers, by the articles they write and the journals they read and publish in, tend towards the description of the problem (such as cancer epidemiology) while others look at the effects of specific solutions (toxicology). In Swales' terms (1990:32), the discourse community has differentiated goals that can be ordered into a hierarchy according to professional and service relationships between the disciplines and the rhetorical relationship to the problem of the disease.

Only five of the informants contacted in this study declared themselves cancer researchers. The other researchers classed themselves as 'structural / medicinal chemists' or 'pharmacologists' (the terms being largely analogous), interested in cancer research as one avenue of application in their fields. This also becomes evident in many of the pure chemistry papers in the corpus- where pharmaceutical applications to high profile diseases such as cancer (but also and at the same time HIV and tuberculosis) are mentioned initially as ultimate justifications before the details of pure chemistry.

However, the idea that there are some researchers 'close to' cancer research with others at the periphery is only a partial picture. The structural chemists in our survey (*SF, BF, JG*) had recently won a substantial grant from the Cancer Research Campaign - yet during the survey they denied that they were involved in cancer research *per se*. Funding is therefore not a clear guide to an individual or group's perception of community, at least as they present themselves to outsiders. In addition, one informant admitted an unofficial policy of understating an individual's involvement in cancer research because of animal rights attacks which the department had recently suffered.

To what extent the individual researchers associate themselves with 'cancer research' or 'chemistry' is therefore a complex issue. In the survey this became an important rhetorical issue of how the researchers justified themselves to a non-initiate outsider. The chemists explained their approach to the problem in terms of combatting disease with targeting 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 the cancer researchers often commission structural analyses on biochemical compounds from the chemists, the two research programmes are 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*).

7.5 The value of pharmaceutical research to linguistics.

As stated in the introduction, a major linguistic motivation for studying pharmaceutical and cancer research is that these fields involve a high degree of abstract pharmaceutical knowledge. The details of some of the principal concepts included in the work of the department are described in the survey, and from this we assemble the *ideographic body of knowledge* as Guba and Lincoln (op. cit) put it, which represents not a conceptual map but a series of interconnected key terms and formulations that allow us to interpret and place in some kind of hierarchic system the rhetorical aims of the research papers in the corpus.

This approach may even be the strategy for coping with terminology adopted by the researchers themselves - none had taken part in writing courses or courses dedicated to terminology or even formal processes of structural chemistry. Generally, most indicated that teaching involves the area of the chemical nomenclature and its structural rules that constituted a specific topic (such as *polymers* or *proteins*) and that this topic may have an adapted or even unique structural system. All also stated that in their reading of other research, most terminology was couched in the text itself: the structural chemists claimed that only the most familiar structures could be recalled and recognised immediately.

Thus chemists depend very much on the text for their acquisition and understanding of terminology, a priority that accords with Godley's (1993) observations of rhetorical value inherent in the variety and ambiguity in chemical nomenclature and graphic representation. In order to demonstrate the interrelatedness of pharmaceutical and cancer research, it is

necessary to outline the basic terminology that is referred to in later sections. The conceptual system that begins to emerge from this gives meaning to much of the professional and phraseological discussion that follows.

7.6 The terminological world of chemistry and cancer research.

The exercise of explaining the basic terms of cancer research is an important first step not only in setting the corpus in its proper context, but also in highlighting the differences of formulation and reformulation between this lay account and the professional use of the terms to be analysed in the corpus. In addition, any problem - solution pattern we might wish to postulate for cancer research (as seen later in the corpus) is complicated by the fact that cancer is not one but many diseases. This complexity can be demonstrated by a brief introduction to the subject using major concepts from the survey that happen also to be high frequency items in the PSC corpus. These terms are indicated in italics in this section. The explanation is based on an expert informant's (*MT*'s) description and on an introduction to the subject by Thomas and Waxman (1991:1-15).

All cancers have in common a genetic virus (promulgated by a potentially malignant part of a gene: the *oncogene*) that produces defects in the ways *cells* are reproduced and developed according to their predetermined function in the metabolism (the undiseased 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. In the department, this complexity accounts for a wide variety of specialisms, going beyond the field of genetics and involving the organic chemistry of compounds that come into contact with the cell. For example, malignancy not only involves *growth factors* (especially *TNF* in our corpus) attaching themselves outside the cell, but also the activation of oncogenes in the cell nucleus where *ras* proteins are able to transform *DNA* within the nucleus itself.

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 by steroids and peptide *growth factors* (complex chemical proteins such as *kinases*) in breast cancer. There is however no consensus on the molecular origin of malignancy (Thomas and Waxman 1991: 6). The only generalisation appears to be that diet is by far the largest cause of growth factor activity, followed by tobacco, viral infection

and environmental influences (such as electronic radiation). There is also marked empirical controversy, since some human tumours are known to be caused by DNA related viruses (for example, immunodeficiency virus is associated with AIDS related tumours) while most scientific research has centred on simpler animal RNA viruses (1991:5).

Because of the uncertain nature of malignancy, pharmaceutical responses to cancer are varied. Generally, intervention in 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. In the pharmaceutical sciences department, there has 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*. Some of the researchers in the department are interested in the chemical transformation of information when the growth factor is chemically *synthesised*. By developing compounds that can target cells and replace receptors or growth factors, a receptor can be developed that finds (*targets*) and destroys (by *inhibition*) the incoming growth factor (or conversely a tumor necrosis factor *TNF* that 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. Some researchers are interested in the process of inhibition itself, others in the possible starvation of the tumor's own metabolic system. Since most researchers are interested in possible treatments of cancers, their specific role in this framework is explained later in the survey.

By introducing the central terminology of cancer research here, the relationships and aims of the pharmacists and molecular biologists of the department can be better envisaged. Two planes emerge. Firstly, research can be situated according to the parts of a cell the researcher is most concerned with, such as the molecular processes within and surrounding the cell. Secondly, specific research can centre on the description of the effects of the disease, or causality and chemical intervention against the disease, in other words according to the complex concept of disease. A researcher cannot be permanently placed along these lines: his or her reading may cut across most of these boundaries, so we can only speak of current concerns, prioritised according to what funding the researcher can get, which areas are prioritised by research teams and colleagues, where most of his or her research has been published and how the researcher presents himself or herself to the lay investigator.

This section serves as an initial view of how scientific meaning in cancer research is structured. How this interrelates with the professional and interactive use of the terms in

research articles is set out below: the survey provides a professional and rhetorical dimension to this picture, while the corpus analysis provides evidence of how these meanings are expanded and negotiated through reformulation and phraseology.

7.7 Details of the Survey.

In order to gather contextual information for the three main hypotheses of this thesis, a questionnaire was prepared and interviews arranged with 15 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 genres in that community (6 questions). All responses were taped, and relevant data and quotations were written down over a period of six months in 1993.

7.8 Survey questions one to four: The discourse community.

An extensive survey of 20 000 academics by Boyer (1994) has shown that researchers have a greater sense of identification with their discipline than with their own institution. Also, as mentioned above in his closer analysis of scientific researchers, Myers (1990) has demonstrated that the field of expertise, expressed through rhetorical stances, reflects professional relationships, internal institutional organisation and the agenda of research and publication. Myers' starting point was that many 'contextual' studies of science (including 'language audits') concentrate only on sociological and institutional data. Thus the ideographic body of knowledge is a more useful key to the discourse community than other sociological factors. By exploiting the expertise of the respondents, then, the following questions elicit terminological information, but also provide us with a topography of how the researchers see themselves in relation to the research specialism.

7.9 Findings from the Survey.

Survey question 1) What is your title and position within the Pharmaceutical Sciences department?

The survey involves: 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*). The institutional affiliations have been noted above.

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

The symmetrical way the scientists fitted 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 chmeical 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. The differing perceptions arise from the complexity of the problem, and from the 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?

Microbiologists and pharmacists were divided 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 should state the aims and methodology of the research discipline, although it is hard to see how these cannot also include claims of centrality and individual originality, and this is how most 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?

Despite the categorisation of researchers into declared specialisms and research groups, none of the researchers said that they worked in formal research teams, despite shared concepts and objectives. Instead, when researchers related their research to other fields they would personify the field as the interest of a colleague, for example, "*RW* would be interested in that." *WF* referred to a "common pool of experience" - in line with the distributed view of the problem of cancer sketched out by survey question 2. Inside the department, seminars and research group meetings were a way of formally discussing findings, while most detailed discussion of research methods and other details took place between individuals and their immediate superiors - either research supervisors or on a very infrequent basis with senior academic staff.

But there were clear areas where researchers' autonomy was restricted, and all of these were linked to the production of written genres. The first of these are official policy documents that declared long term common research programmes which ultimately determine renewal of fixed term contracts. Also, because of the amount of data many papers are split into several sections and published as a series (*SF* was on part 7) and this committed researchers to continue a series of long term research articles. Finally there are grant proposals, obliged by the funding structure of the department and written and submitted by a team.

Outside the university, *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, *WF* suggested that there might be around 10 people in the world who might be considered experts on his own specialist compound. *AG* also noted that researchers would be aware of related groups which would be regarded as 'soft competitors' exchanging research papers and communications, coordinating some grant proposals, at other times competing for them. While the cancer researchers saw most of their links with national charities where their research was coordinated nationally, 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?

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. 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 by commenting on which articles would attract their attention: *JG* proposed that he read around 10 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. Researchers stated that specialist entities (a term we 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:

> What things does it deal with?
> Has anyone done this before?
> Are there surprising results?
> Do I believe it or not?

According to *WI* these would then lead on to specific areas of the journal. In *MT*'s case, surprising results may be due to the number of animals used in the study and other methodological details. *PL* also 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. The Journal of the Chemical Society's instructions for authors (1993: xii) 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 centring in by keywords 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 actual 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) is 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 (e.g. cancer) and their phraseological or terminological synonyms (oncology), followed by subject specific co-terms (such as advanced, anorexia, associated, clinical) and then semi-stop procedural words (such as methods, analysis) which 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 the redundancy of high frequency items in indexing. However, the possibility of high frequency items being associated with rhetorical and phraseological patterns in the corpus may reveal new avenues of indexing that exploit such formulations (Ref: Science citation index 1993 *Permuterm's* list of semi-stop and stop-words).

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

From question 6, two reading patterns appear to be browsing and consulting. What information can be derived from different parts of the article therefore depends on the expectations and expertise of the researchers. The more experienced researchers may 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 titles were 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, or when the author's comments are not supported 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 finds 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 non-reading and partial reading are not indicative of irrelevance but simply of the researcher's confidence in imposing coherence independently from the text.

The kinds 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.
4) and ('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 'given' 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 in the abstract. *DL* stated that the discussion section evaluated the current research, as well as suggesting or predicting an extension to the research model.

The pharmaceutical scientists (*SF, WF*) said that there was an overlap between the methods and results sections, since methods sections start off as lab book transcriptions combining a template of measurements. This corresponds with an unexpected symmetry in the corpus: all of the experimental sections in the corpus occurred in chemistry journals, and these often replace methods and results sections in these journals (especially the shorter communications). In contrast, the microbiologists (*PL, MT*) saw results and discussion sections as distinct from methods, and in the corpus all the amalgamated 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 actual results should be interpreted and integrated in the context of medical applications. Presumably experimental data (or integrated methods and results) for the pharmacists can stand alone, such that the shape of the data and medical applications can be treated separately in the discussion section. This implied distinction between applied biochemistry and theoretical chemistry may be an oversimplification, but any phraseological distinctions between these two essentially rhetorical positions can be elucidated in the corpus analysis.

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, essays such as TPS, 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). This question was accompanied by a request to donate a published research paper for use in the corpus. This variety is an important consideration in the selection of texts for the corpus, and is discussed in Chapter 8.

Survey question 9) Details of writing up.

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

MT admitted 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 he termed 'stories') 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 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 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* indicated, 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 retrieved and read in non-linear fashion, their production appears to be just as non-linear, essentially being built and redrafted by several writers from the 'middle' out towards the introduction and discussion sections. Different members of the research team record reaction details of syntheses and other measurements over a period of months in the lab book with its various sections:

-Title (of extreme importance to avoid confusion of data)
-Date (to avoid repetition and to measure stages of progress)
-Reaction name
-Structural formulae (materials involved listed in shorthand codes)
-Reagents (catalysts and added materials for synthesis)
-Procedure
-Structural analysis of final product (in molecular percentages)
-Specific measurement details:
(yield, melting point, optical rotation, refractive index, elemental analysis...)
-Purity (checking contamination)
-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 forms the backbone of the research article that can be fleshed out by adding explanations of unfamiliar procedures. The computer can be used to automatically create tables and the researchers mentioned programs that would check the locations of references to tables and figures within the body of the text, as well as drawing chemical formulae automatically from structural names (very often codes). These practices

account for the number of mostly unrecoverable abbreviations and long lists within the corpus (leading to the elimination of many non-linguistic experimental sections), supporting Swales' (1990) comments that methods sections do not encourage replicability or accessibility.

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

All the researchers referred to specific journals' instructions for authors. 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 as chemical bonds) 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

> i) originality of scientific content and
> ii) appropriateness of the length and quality to content of new science.
> (1993:vii)

When asked what changes referees require, MT stated that they generally correct structural aspects of papers, tone down claims and question the generalisability of experimental data. Other researchers had many examples of correction of style, *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 an excuse, if charitable". *PRL* mentioned stereotypical phrases such as "typical results show that" and "preliminary experiments have shown that". Ironically, we demonstrate in Chapter 11 of the Data analysis that these are some of the most frequent and consistent expressions in the corpus.

8.0 CHAPTER EIGHT: The Corpus.

The survey in Chapter 7 has detailed how research activities are organised in the pharmaceutical sciences department, how research problems and solutions are perceived and what constitutes a new claim in the eyes of the discourse community. In this chapter, a corpus based method of data collection is proposed to address the reformulation and phraseological hypotheses. The chapter then justifies the choice of texts for the corpus in terms of how these relate to the discourse community, and describes the typological characteristics of the main Pharmaceutical Sciences Corpus (PSC) with its associated control corpora.

8.1 Aims of the corpus.

> 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)

The phraseological hypothesis and its corresponding corpus analysis constitute the central methodological objectives of this thesis. The reformulation hypothesis explores data from the same corpus, while the reconceptualisation hypothesis motivates comparison between the PSC and the other control corpora. Before turning to corpus design, it is necessary to restate explicitly how the aims of this thesis (expressed by the three hypotheses) are to be achieved, and how the corpus is to be shaped in order to satisfy these aims.

8.11 The corpus and the reformulation hypothesis.

It has already been seen that Halliday's (1993) analysis of the progression of grammatical metaphor in a technical text requires manual, linear analysis of the progression of a concept in a text. Hoey's (1991) and Phillips' (1985) exposition of repetition in text provides a method of selecting a recurrent concept throughout the text, by selecting sentences bonded with three or more repetitions, synonyms or (importantly) paraphrases. These topics have been discussed in the literature review, and methodological detail which goes hand in hand with analytical problems is discussed in the analysis section.

A large number of texts were not considered necessary for the study of reformulation since local patterns of reformulation in part of the corpus should be relatable to large scale patterns of phraseology which are identified automatically. Since the survey respondents submitted ten texts, these would be sufficient to gather information on the possible permutations of expression that are available in these texts. This presents three advantages. Firstly the texts were included in the PSC corpus and could be considered a representative sample since they cover cancer research, microbiology and pharmaceutical sciences. Secondly, respondents were easily contacted for feedback on the analysis and clarification of ambiguous or unknown terms. Finally, *MT* submitted three texts that were on the same topic but of different theoretical levels while *SF* was involved with the writing of three texts (two submitted by other respondents) which range from communications to fuller research articles. The sample of ten texts could therefore be considered to represent texts that have a topical coherence but differ in coverage, audience and rhetorical intent. A better understanding of differences in this area would make up for the fact that the PSC corpus is differentiated particularly deeply in types of journal and topics covered and less deeply in terms of small differences of genre.

8.12 The corpus and the phraseology hypothesis.

For the most part, the phraseological analysis of this thesis is concerned with a large scale characterisation of the corpus that ignores the linear and thematic development of phraseology within the running text, concentrating instead on the recognition of regular patterns and identification of rhetorically motivated deviations from them. This contrasts with the 'linear' analysis of the reformulation hypothesis.

According to the phraseology hypothesis, the key to understanding the role of phraseological patterns in linear progression may lie in identifying patterns within specific rhetorical sections of the research article. This constitutes the main analytical activity in the thesis: characterisation of the phraseological patterns of rhetorical sections (Title, Abstract, Methods, Results, Discussion) compared with the corpus as a whole. However, it has already been seen that different specialisms may have different perceptions of how experimental, methods, results and discussions sections interrelate. Corpus design must take this into account: either by the selection of representative texts where the rhetorical sections have an equivalent role or by interpreting findings according to research specialism, type of journal and other contextual factors: such as the omission of one section in a slightly different genre, as in the short communication or essay. Also, in order to be

properly represented, the smaller rhetorical sections (Titles and Abstracts) must also be compensated by control corpora, and the selection of these must accord with the selection criteria of the original corpus. These questions are addressed in the next section.

8.2 Corpus Design.

Having explored the main objectives of the corpus as they correspond to the research hypotheses, it is now necessary to set out the principles underlying the choice of texts for the Pharmaceutical Sciences Corpus. The methodological advantages of corpus analysis for a description of languages for specific purposes have been set out in the final chapters of the literature review. In short, it became clear that a relatively small corpus would be large enough for analysis of a specialised area. Secondly the rhetorical aims of the writers had to be prioritised in the analysis. This consideration was not the primary aim of the original Cobuild project. For example, Renouf (1987) describes the texts used in the Cobuild corpus in terms of very broad categories ranging from broad registers (non-fiction, procedures, argument-positional, narrative) to specific genres (surveys, the NATO-corpus, the Sizewell enquiry corpus).

However, as computational corpus-based translation, terminology and lexicography diversify, Sinclair as well as others such as Atkins, Clear and Ostler (1992) and Ahmad et al. (1991) have argued for a greater contextualisation of corpora. In this perspective, Sinclair (1993c:6-7) proposes four principles of corpus design to which the following four sections of this chapter correspond:

1)The choice of texts should be governed by a stated view of language in communication.
2)The variables determining the choice 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 are concomitant with principle 1.

8.21 The language view of the PSC.

As stated earlier, the research article, in its already diverse forms, is seen as a privileged statement of 'public' research and is thus the main object of enquiry. Other texts, such as grant proposals and internal documents of the department, 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 the genres of the discourse community therefore, a rhetorical overview of the department can emerge from a mixture of authors' own texts, texts that are

considered to be central to the researchers' work, and texts that appear in the journals they regularly read. Criteria for choosing these are set out in the next section, the aim being to create a corpus that is coherent given the limitations that we have set out here. Even taking into account the differences between researchers and specialisms represented in the corpus, few other corpora have been analysed lexico-grammatically at a similar degree of specificity. By delimiting the rhetorical, topical and generic variables as far as possible, and signalling where there are wide gaps between variables, phraseological deviations can be measured against a norm and associated with probable rhetorical aims corresponding to those set out in the survey.

8.22 Conditions of inclusion in the PSC.

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

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:

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.

Such variables cannot be made entirely distinct as Sinclair (1993c:6-7) may have wanted them. In the PSC corpus, a combination of these criteria can be seen to be operating, and some criteria account for more research articles in the corpus than others (especially *author, prestige* and *centrality* but also *access*: see below).

All of the fifteen 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 heterogenous body of texts, not only in terms of different sub-fields as mentioned above, but in 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 a wider than average readership: *Trends in Pharmaceutical Sciences* (TPS), another gave an article in *Tetrahedron Letters* (TL) which is an incomplete part of a series of communications on a specialised drug where the readership would be highly limited.

One solution might be to calibrate the papers in the corpus 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 mentioned earlier, the department judges its own research publications according to the Science Citation Index impact factor. While papers in research selectivity exercises are strictly 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 Index, while the well known high-circulation journal *Nature* (14th) is preceded by the esoteric *Advanced Cyclic Nucleic Proteins* (8th) (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. This issue is particularly relevant to the building of the corpus: it is not enough to state that a corpus represents 'prestigious journals in the field' where even an objective measure attempts to distinguish this. Nevertheless, the measure does have some importance, since it is valued by the institution and external funding councils, if not by the individual scientists themselves.

Also, during the survey, some researchers were keen to point out the relative values of some journals over others: *Tetrahedron Letters* was of doubtful quality according to another researcher (*DP*), because it publishes communications which have not had time to be tested, or in Myers' words, to become accepted ideology. One way around this problem is to ask experts for journals they had been using for reference as well for journals they thought of as key to their work. While looking at texts that the researchers value for their own particular needs, this ensures that the research papers represent the wide range of journals and topics that the researchers must read or be aware of but do not necessarily publish in.

This was the system used in the survey, although as seen below, it failed to qualify the JGM as a *prestige* journal in *DP's* term but simply as a specialist *central* one.

8.23 Documentation of the PSC and Control Corpora.

8.231 Choice of Material : PSC

The compilation of the PSC corpus involved collecting research articles from a selection of journals, optically scanning and storing them on floppy and hard discs. Although the decision to use each journal is motivated largely by the contextual reasons given below, specific papers from these journals were obtained largely at random and were expected to represent a large readership according to the 'grapeshot principle': a large number of texts would represent and interest a wider group at the same time as being typical of the type of text read in these journals rather than being the actual texts read by individuals. The number of articles collected from each journal was largely determined by how many papers were available for each journal, copyright restrictions (such as not copying the whole journal), length of article, and quality of paper for scanning. The following working conditions of inclusion in the corpus emerged as the survey continued and as the need for texts arose, and correspond to a rough combination of the criteria set out above:

1-*Authorial*: The corpus should include any research articles the researcher had (co-) authored. Ten articles were obtained this way. One researcher submitted three papers, another two papers (one in electronic form) and five others submitted one each (one in electronic form). Five researchers did not donate an article.

2-*Prestige*: The corpus should include any research articles from journals that researchers mention more than twice in survey question 5a (articles the researcher has recently read to catch up on his or her research field). This accounted for 80 of the 150 research articles in the corpus.

3-*Centrality*: The corpus should include research articles from journals mentioned in survey question 5b (journals the researcher has recently read and considers key to his or her research field). 36 articles were obtained from the ADONIS biochemistry on-line catalogue.

4- *Accessibility*: FAT, JPP and CAR were available on *Medline* and could be immediately downloaded. Article AC was submitted by a researcher from Birmingham University who

didn't take part in the survey. This gave 24 articles.

The numbers of articles per journal and the main reason why the journal was obtained are noted below according to journal code as listed in the PSC Reference Lists (Appendix A).

Motivation of Choice of Articles with Numbers of Papers.

By author:	BJ, CC, JCPT[7, 8, 9, 10], JMC, JNCI, TL, TPS
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]
By topic centrality:	BJC[1-11], CL[1-9], JGM[1-9], JOC[1-7]
By accessibility:	AC, CAR[1-10], FAT[1-10], JPP[1-3]

In Appendix A the corpus is documented in terms of Journal SCI Rank, percentage size of the corpus per journal and title of each research article. The rhetorical, topical and textual breakdown of the texts are also detailed below in section 8.24 on the constitution of the corpus.

8.232 Choice of Material: The Control Corpora.

It was decided that the PSC corpus would be split not only into topical sections (pharmacy and cancer) but also into rhetorical sections. Two rhetorical sections needed augmenting to strengthen statistical characterisation of TAIMRD sections to provide more examples of phraseology: titles and abstracts.

Although the original 150 titles and abstracts of the PSC corpus are compared with other rhetorical sections, a subcorpus was derived from the freely accessible electronic index, *Medline*. The *PSC-Medline* subcorpus consists of the first 572 abstracts (a convenient discfull of 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 journals (foreign language abstracts were discarded) and relate to cancer either from within the title or abstract or from the list of keywords included as *Medline* data (but the keywords are also discarded for this study). The *Medline* corpus thus has the advantage of topical specificity as well as a homogenous genre. In the data analysis section, we compare the PSC titles subcorpus with the PSC corpus as a whole to give a picture of the

salient lexical items which are typical of titles with the PSC corpus. These results can then be analysed using the *Medline* corpus, since the PSC titles corpus alone is not large enough to be reveal interesting concordance data.

8.233 Preparation of Material: Practical considerations.

When the 80 'prestige' texts had been chosen for the PSC, they were scanned page by page using an Optical Recognition Device (ORD) linked to an Apple Macintosh. Copyright restrictions meant that only one article per journal could be legitimately copied. The problem posed by the large number of journals this procedure required was partly overcome by borrowing journals from the researchers themselves, rather than the university library. This had the added advantage that the journals could be said to be 'lab copies'. However, because of the complex nature of the field and the sometimes poor paper quality of journals or photocopies of research articles, some typographical errors still remain in the corpus. A particular problem that accounts for certain anomalies of word counts is the number of scanning mistakes due to small print. In many cases, this meant that experimental sections had to be discarded. The texts that accompany tables have been included at the end of sections which refer to them. Once post-edited, all the texts were converted to text files for use on a PC mounted UNIX system for frequency tests and then ASCII files for the PC mounted MS-DOS concordancer (detailed below).

8.24 Constitution of the Pharmaceutical Sciences Corpus.

8.241 Textual Overview of the PSC.

The PSC corpus consists of 150 research articles. Using Roe's methodology (1993:10) a UNIX word frequency count (taking a word to consist of any string of symbols bound by two spaces, excluding figures) calculates the total word count to be 515 073 running words (tokens). The number of words is probably exaggerated (there are chemical and Greek symbols that may have inflated this number). A second count by the Wordlist program (Scott 1993) gives 499 105 words, of which 24 253 are different words or types.

The PSC corpus is split into rhetorical sections (subgenres if abstracts are considered as such) of which UNIX wordcount calculates the following proportions (adjusted to take account of overlapping such that percentages are percentages only of other texts that share the same format):

Subgenre Total	*Tokens*	*% of PSC corpus.*
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]	

Hybrid rhetorical sections which replace the function of two separate sections occur in a portion of the corpus (Methods/Results, Results/Discussion). They have a percentage based on the total number of texts that only have those sections (henceforth termed MR-sections, RD-sections) and although these figures suggest they are large sections, they are proportionally smaller than the corresponding non-hybrid sections when these are combined. There are hybrid rhetorical sections in 30 articles as well as 9 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 sections. 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 the element of overlapping (for example, some articles having RD and E-sections such as the structural chemistry JCPT) 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). RD-sections however do replace separate Results and Discussion sections (including subsumed Conclusions and a Summary). For the phraseology hypothesis, however, we look only at T A I M R and D sections, bearing in mind that the control corpus would be 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 do rate very highly in a list of 8 000 journals, but not necessarily according to the classification obtained from the survey ('prestigious' journals are underlined for comparison):

Table 1 SCI Impact Ratings of the PSC Journals.

Journal Name	SCI Rank (1988)	Journal	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 within the first 600.]

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. In terms of relating the PSC corpus with its discourse community, PSC includes many high impact journals, and has quite a specialised coverage with the exception of such 'introductory' articles as TPS and those in the *BMJ*.

8.242 Topical Overview of the PSC.

Having compiled the PSC corpus, the next stage involves checking the topics of the papers with the researchers to determine the exact fields involved in each paper. Two researchers (one from each main division) helped to classify and gloss all the research articles in the PSC according to the following research specialisms:

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 thus has a strong cancer research bias (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 pharmaceutical sciences part of the corpus (63 articles) is more general, covering perhaps a fraction of the diverse specialisms of the field, with very general fields such as 'structural chemistry'. The pharmaceutical side can be seen very much as a 'satellite' topic with indirect links, as set out in the survey, with the problem of cancer research and disease in general. 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). No examples of BMJ papers on cancer were available, so five random papers were included as examples of the genre.

8.3 Corpus Typology.

Having set out the basic internal structure of the PSC corpus, it is now possible to establish how the corpus (and its control and subcorpora) compares to other corpora before the analytical methodology and the analysis of the corpus can take place. This is essentially motivated by the desire to account for context in the corpus at the same time as providing data for comparative studies to judge any findings in relation to their corpora.

Atkins, Clear and Ostler (1992) have set out a taxonomy of corpora for the description of the International Corpus of English. In terms of their corpus typology (1992:13-14), the PSC corpus is a monolingual, composite full-text corpus. Although terminological data can be gathered on the specific area from the corpus, the corpus is not tagged for terminological use and thus the PSC is for 'general' use. Atkins et al. consider a corpus which has a 'shell' representing the rest of English for Technical Purposes and a 'core' representing the commonality of the language. The PSC can be considered to be its own 'central' corpus and this only in relation to the *PSC-Medline* subcorpus of subject specific (cancer) titles and abstracts. However, at the beginning of the characterisation of the corpus a statistical comparison is made between both the PSC and the original Cobuild corpus (17 million words) to determine general lexical differences as though the Cobuild corpus were a 'core corpus' applicable to the general language rather than for specific purposes.

8.4 Text Typology.

Atkins et al. (1992:15-19) also propose a typological template to establish the various biases of the corpus. They reject the idea of an intuitive 'balanced' corpus on the basis of extra-linguistic features, if only because even if criteria could be agreed, it is impossible to impose balance at the beginning of the corpus building process. Enkvist (1964) provided an early characterisation of context that has become a model for corpus linguistics. Enkvist differentiates between textual context and extratextual context. Textual context includes the traditional linguistic levels together with the format and outer appearance of the text. Extratextual context consists of the period, type of speech, literary genre, participant relationship, sex, age, class, status, common stock of experience, context of situation, and

physical environment.

In terms of Atkins et al.'s typology, the PSC is a written-to-be-read, multi-person, prepared set of periodical texts:

- PSC **'style'** is clearly 'academic scientific' and presumably varies according to internal factors such as *coverage.*
- The **'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. Elsewhere we have occasionally use the term co-genre for these, and rhetorical section or subgenre for such texts as 'introductions'.
- **'Function'** would be an oversimple category: although the PSC may be covered by such terms as 'informative, persuasive' rather than instructional.
- As for **'setting'**, while belonging institutionally to education, the PSC clearly is based on a 'scientific research' setting, including laboratory use with the other contexts where the research articles are consulted.
- **'Topic'** has been set out above, although the statement "This text is about X" with possible responses "science, biology, chemistry. etc." Atkins et al. (1992:17) reveals the difficulty of separating topic from *coverage.*
- **'Technicality'** is defined as: "based on degree of specialist/technical knowledge of the author and target readership/audience" and is an external variable: the PSC is technical as opposed to semi-technical or general.

Details of 'authorship' such as 'authority' are only known for the texts originating from the survey, and other factors are largely unknown: many of the names indicate that the majority of the texts in the corpus are written by teams of non-native speakers. Despite the large number of multiauthor texts, there is no evidence to suggest that single-authorship is indicative of topic coverage or authorial authority: single author papers AC and TL are very specific and written by post doctoral research fellows, CC is a specialist specific single author text by a senior lecturer, and TPS is a more general text by a professor.

8.5 Text Analysis.

The preparation and compilation of the PSC corpus is essentially where methodology and analysis meet: by the procedures adopted to characterise the PSC texts an implicit analytical scheme is already being mapped out. In this chapter main analytical procedures are detailed in order to leave the linguistic analysis proper to the data analysis section of the thesis (Part III: Chapter 11). Here type / token frequency, comparison with Cobuild frequency lists, statistical comparison of different rhetorical sections and other basic computational procedures are set out.

The procedure used to prepare and compile the PSC corpus is similar to that used in the compilation of the Cobuild dictionary (Krishnamurthy 1987, Clear 1987) and has been broken down into a series of computational steps by Roe (1993:10-13) on a UNIX based system called the *ASTEC* suite and later developed for the WINDOWS environment as the *Aston Text Analyzer* (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 later by an MS-DOS based collocation program (*Microconcord*: Johns and Scott 1993) and the WINDOWS-based wordlist compiler (*Wordlist*: Scott 1993). The differences in operating system for these various programs, their varying definitions of what is an acceptable and unacceptable 'token', and textual changes of format in converting the PSC corpus for these systems mean that consequent differences in word frequency lists must be taken into account.

8.51 Stage 1: Analysing frequency.

The main justification for frequency lists in this thesis is the capacity of the computer to statistically indicate the most salient lexical differences between two texts or corpora, thus indicating in a replicable way where the analysis of phraseology can take place. In *ASTEC* a frequency list is first created for the whole corpus. The definition of a word or token according to *ASTEC* is "a sequence of letters, bounded by spaces, within which the sequence letter/hyphen/letter and/or the sequence letter/apostrophe/letter may occur more than once." (Roe 1993:5). *Wordlist* gives the following percentages of the PSC corpus for

the first 10 items in Cobuild and in the PSC for comparison:

Table 2: The Wordlist top ten lexical items in the PSC and Cobuild corpora.

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

Already clear differences between the highly specialised corpus and the general corpus can be seen, especially in the sharp increase in the proportion of prepositions other than *to* in the PSC which have been found to be indicators of interesting differences in phraseological patterns (Gledhill 1995). It is interesting to note differences with the original Cobuild list (Roe 1993, Sinclair 1991) such as the drop of *that* at rank 7 (rank 12 in the PSC with 3 359 occurrences), and the loss of *it* at rank 8 in Cobuild (but down to rank 41 in PSC with 1 006 occurences). Also the decrease in the use of the indefinite article is noticeable, even with the possible increase in the use of the letter in other contexts in the PSC corpus.

The *ASTEC* frequency list gives a more global view of these figures by comparing the relative frequencies of the PSC corpus out of 10 000 with a similar 17 million word Cobuild list. This gives two specific results. The 'COMMON' list contains a list in descending order of relative frequency (per thousand: effectively the equivalent of a percentage) of each item in the PSC and a figure indicating the relative frequency in the Cobuild list. Further down the list a clear pattern emerges: clumps of words are very significantly associated with the PSC corpus (*between, human, table, using, results, both, study, shown, protein, observed, DNA, data* are all at 0.14% or more compared to their occurrence in Cobuild: 0.7% or less. Other words are biased to the PSC, but not as significantly: of which the most frequent are: *of, and, in, was, with, for, were, by, cells* (0.6% versus 0%), *at, from, or, et al., these, after, also, mice, activity*. Conversely, Cobuild- oriented words include: *the, a, to, that, is, as, on, this, are, be, not, which, an, have, it, all, has, but, other*. After these items the list clearly separates PSC oriented words (with high percentages on the left) from the Cobuild oriented words (higher percentages on the right). While this list contains the most common items, the PSC 'Salient' list contains

the first 100 items that occur most frequently in the PSC corpus relative to their occurrence in Cobuild. The PSC salient items: **of, and, in, was, with, for, were, by, after** are studied in particular detail because they are salient to particular subgenres. The items *the, to, that, is, this, be, not, have, all, has, but* are also salient to subgenres - but it can be seen from Table 2 that they are globally more typical of Cobuild than of the corpus in general. We expect this to say more about the relation between the rhetorical section and the rest of the corpus than between the rhetorical section and Cobuild. Since the type of analysis these lists allow is restricted by the rhetorical and topical distance between the Cobuild and PSC corpora, the results are taken into account as secondary evidence for the main methodological impetus of the thesis: the comparison of rhetorical sections. For some very high frequency words we compare patterns in the PSC with the Cobuild 1995 dictionary as an indicator of 'the general language'.

8.52 Stage 2: Determining salient items.

A salient item is a word that occurs significantly more in one text (or part of a text) than it does in another. Using the *Wordlist* program we select ten of the most salient items from each subcorpus in order to examine their phraseology. Before describing how this done, we need to emphasise here that for maximum coverage we select the ten most salient *grammatical* items from each subcorpus wordlist. The rationale for not studying simply the *most salient* items in any lists is set out later in section 8.6 and again in Part III: data analysis. We also signal here that all types of Methods, Methods/Results and Experimental sections are combined for the purposes of analysis but Results-Discussion sections are kept separate from the Results subcorpus and the Discussion subcorpus. Results-Discussion sections are taken into account in the statistics for the whole corpus but are not the subject of phraseological analysis in this thesis. 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.

The *Wordlist* program compares proportional lists made for each rhetorical section of the PSC corpus, weighing the frequency of words in each list against the proportion of the corpus made up by the subgenre, and prepares a list of salient items for that rhetorical section. According to *Wordlist* here are the proportionate sizes of the most important rhetorical sections:

Table 3. Subcorpora compared in the *Wordlist* analysis.

Subcorpus	*% of PSC*	*Tokens*
Title	0.5%	2 127
Abstract	5.8%	29 136
Introduction	11.8%	59 724
Methods	27.5%	137 161 [=M+ MR+ E]
Results*	25.8%	119 746
[Results-Discussion*	46.1%	36 647]
Discussion*	24.8%	114 829
Total PSC	*100%*	*499 370*

(* % adjusted to compare only with texts of similar format).

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) giving a chi-square score of significant difference (as described by Butler 1985:176). 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 PSC. The results of the most statistically significant salient items for each rhetorical section (termed title-, abstract- salient items etc.) are listed in Appendix C. We have only listed the first and last pages of these: a wordlist comparison compares every word and these lists are too long to be included even in the Appendices. To demonstrate the use of these word lists, here is a selection from the most abstract-salient items in the corpus:

Table 4: from *Wordlist* : Abstract-salient items in the PSC.

Rank	Word	Freq.in Abstract	PSC Freq.(%)	Chi^2	p
31	but	67	(0.2%) 663	18.1	0.000
32	immortalized	13	(0%) 69	17.9	
33	showed	43	(0.1%) 375	17.4	0.000
34	increased	43	(0.1%) 376	17.2	0.000
35	interval	12	(0%) 56	16.9	

Near the bottom of the list in Appendix C, *immortalized* is the 32nd most abstract-salient item (by virtue of its observed frequency in the abstract (13) divided by its observed frequency in the PSC (69)). But its occurrence is not statistically significant when the size of the abstract corpus is compared with the expected frequency in the abstract (the chi-

square is shown but as it is not significant the p score is not shown by *Wordlist*). On the other hand, ***but*** is the 31st most abstract-salient item, the first grammatical item on the list and has a chi-square score of 18.1, which at 1 degree of difference (Butler 1985:176) places it even below the 0.1% level of 'very highly significant' (5% is usually regarded as significant, and those items with a $p = 0.000$ score in the lists are thus considered very highly significant). *Wordlist* also 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%). In fact, the occurrence of *immortalized* (13 out of 69) is significant, as is the occurrence of *interval* (12 out of 56) but from the Wordlist tables it can be seen that there is a statistical cut off point in terms of items that are too 'small' to count compared to items from the whole corpus: for abstracts it is 90. This means that while items with less than 90 occurrences in the PSC may be abstract-salient, they are not given a p-score. The last page of the *Wordlist* results is also included for each subcorpus because this indicates the items that either do not occur in the particular subcorpus or are very significantly atypical of that subcorpus. These data are considered in Part IV: Conclusions.

As internal measurements of the relative distribution of words in the corpus the *Wordlist* results serve as the basis for deciding which phraseological patterns are to be analysed. 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 throughout the corpus, but a different phraseological pattern in which case one can only hope to find the different patterns through analysis of the other words. Generally, comparison with the Cobuild list should indicate which items have some role to play in the corpus (by their presence, or by another phraseological pattern in their absence) as well as indicating whether items that do not have a significant role to play within the corpus are generally also constant outside the corpus.

The subcorpora-salient items that emerge from the Wordlist analysis are set out in Part III, the data analysis section. This is partly because the types of item chosen were grammatical items rather than the most salient items. The rationale for this is set out later. Here we list the grammatical items that emerge as salient items, indicating by code their original subcorpus (some items, like 'both' or 'this' are listed in their most frequent word class as observed in the corpus):

Auxiliary 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),

Adverbs (2): then (M), not (R, D)

Of 57 possible items, some words' salience in different subcorpora means that the analysis covers 38 items. This allows for some degree of analysis of phraseological distribution across the corpus: the behaviour of 'in' for example, can be analysed in titles, abstracts and results and discussion sections.

8.53 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 token in the text. The next step involves an alphabetical sorting of the index file. Patterns 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 searching for patterns by, for example, changing the concordance so that words to the left or the right are listed alphabetically. The advantage of *Microconcord* is that patterns are outlined in colour and can be arranged so that the listing is alphabetic for example for two places to the left and then three to the right, highlighting patterns over a long range and permitting the analysis and sorting of collocational frameworks (Renouf and Sinclair 1991). Here is an example from an ordered concordance of the word *of* elicited from the *Medline title corpus* where the left hand pattern was revealed first; then an ordered listing is elicited for one word to the right:

Table 5: Selection from an ordered concordance of *of*.

Anesthetic management	of a patient with Bartter's syndrome.
The neurosurgical management	of brain metastasis from colorectal
Psychological management	of breast cancer patients in a group.
ort review. 371 Management	of chemotherapy-induced neutropenic se
Teicoplanin in the Management	of Febrile Episodes in Neutropenic
Ch resistance in the management	of head and neck cancer.
ent trends in the management	of invasive bladder cancer.
urrent 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.

Clearly, 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). It constitutes a 'expression' because regular topical patterns and deviations from the pattern can be seen around it: firstly it allows expression of general approaches: *current trends in, diagnosis and...* as well as histochemical approaches: *Treicoplanin in, irradiation in, resistance in* and the expression allows precise modification which signals methodology: *anesthetic, neurosurgical, psychological.* 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.

This kind of analysis forms the basic methodology of this thesis, particularly Chapter 11. The advantage of this 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 (*X cancer*, patients). In particular, concordance analysis captures groupings of low frequency items or items which would not be automatically associated with the keywords such as *dysfunction, pain, metastasis, neutropenic secretion.*

To signal where an intuitive reading of the concordance has revealed a pattern, a semantic covering term in brackets (disease Y) is used. Frequently observed collocations are underlined (management **of**), indicating that they are linked with no intervening elements (and bold items indicating PSC salient items) or < + >, indicating that they are statistical collocates but may have other elements between them: management **of** + (cancer). Optional elements can be indicated by a < / >. Thus the phraseology revealed above can be denotated

by the three line formula:

Left Context: (general empirical approach) / (specific empirical approach) /(treatment-related drug X) in the

Collocates: management **of**

Right Context: (part of body) (cancer) / (disease Y) (symptom)

In the phraseological analysis section of the thesis I have used four major semantic categories: research, clinical, empirical and biochemical, with certain further subcategories. I have used the symbol X to demonstrate the many types of treatment-related names of compounds and Y for many disease-related items. In order to make examples as accessible as possible, an optimum of five concordance lines is shown for each pattern identified.

8.54 Stage 4: Calculating collocation.

All words that co-occur within an arbitrarily determined distance or *span* of the node are termed collocates, and collocation is seen as the frequency with which collocates co-occur with one node relative to their frequency of collocation with other nodes. All that separates collocation from mere word cooccurence is the statistical level at which the researcher is happy to say that the cooccurrence is not accidental. Sinclair (1991:68) exemplifies 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'* plus 556 instances of *'off'* in 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 (1987b:153).

For our 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 statistical justification for the assignment of phraseological patterns. The term 'phraseological collocation' is used to signify patterns that are not significant or even frequent by themselves but are visibly (or intuitively) part of a pattern, such as the pattern (empirical process) in the management of

+(disease Y).

A built-in assumption of statistical 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 cotext (left or right of a node) that should be taken into account, on the grounds that if, as Sinclair demonstrated, collocates are not independent variables then there should be some systematic approach to determining statistical dependence. Generally, phraseological methodologies 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. The collocation programs used in this thesis provide a range of means of calculating frequency of collocation (to a span of ten) and position of collocation (to a span of three):

Microconcord: Short range (3 x 3) globalised collocation (either informational or directional)

Astec: Short range (3 x 3) localised collocation (directional only)

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 corpus is:

the (174) a (134) the (574) <u>of</u> the (354) of (67) a (34)

This is useful for determining whether some collocates are distributed according to position and further programs allow for a distributional analysis across several texts (the UNIX *DIST* program), but this does not give an immediate pattern that can be followed up by closer analysis of the 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 cotext, 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

Some patterns appear to be established even across such a wide span (*of +breast, of +human*) which are intrinsically interesting. The program also allows for a distributional analysis not across several texts but within a text, giving a 'bar code' of the cooccurrence of up to three items. However, despite the intuitive satisfaction of accounting for a large area of collocation, other significant patterns that are derived from concordance analysis such as *in the management of* are obscured by the wider span. Work is currently in progress on programs that may take account of this for larger spans (such as Scott's *Wordsmith,* (personal communication)). For all the collocation programs, therefore, it is necessary to establish a method of determining which collocates are significant and which significant collocations render the most interesting phraseological patterns.

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 is actually found and how much is lost. Clear (1993:288) argues that while *precision* is best illustrated by the mutual information (MI) score, *recall* is best illustrated by the t-test. Phillips (1985) and Smadja (1993a) aim at a total collocational description of a corpus, and thus *recall* is an important concept to them. For the purposes of this thesis, 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 cooccurrence 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 patterns that are interesting in themselves: it is only until the bottom 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 research-activity or empirically oriented words like *presentation, department, majority, measurement* which are then qualified by a larger less significant group of semantically more concrete items like (*disease Y, cancer X, patient...*). All of this, of course, can only be related to the corpus in question: thus collocational results are assumed to be characteristics of, in this case, titles in a cancer research index. In fact, 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.

In addition, Smadja (1993a, 1993b) has argued that such analyses are statistically flawed, and his own analysis calculates different positional values of collocation for a span of 10 x 10. While the statistical relevance of this is not in doubt, it has to be noted that Smadja and others' collocational analyses aim at total collocational analysis without relating collocation

to rhetorical usage and thus have a different methodological emphasis than the programmers of *Microconcord* and *Wordlist*. Also, very low frequency items that may have a common semantic relation in a pattern would still not be identified by Smadja's 'total' collocational analysis. For this reason, collocation is seen as essentially phraseological in this thesis, while statistical collocation is a methodological indicator of more complex patterns which require contextual analysis.

For a number of reasons the MI score is not used in this thesis. To begin with, we used fifty collocations of 'of' to arrive at the above table. If we analyse ten items from each rhetorical section, we would have to calculate a large number of collocates for each of the 60 items: that means 3000 (60 x 50) two word combinations. But we are interested in longer collocational patterns than 2 words, and this is not mathematically accurate with the MI test. Another problem of relying on collocational counts is that some word usages (such as the statistically significant use of *but* in the abstract) are significant yet have few short range collocational properties. 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*. We are lucky in that with a relatively small corpus, all of the occurrences of an item such as 'of' can be analysed. The fact the even the highest frequency items in the corpus display remarkably stable collocational properties makes this task easier. To summarise: in this thesis, frequency and concordance analysis take precedence over collocational lists, and significantly salient words (such as *but*) are analysed according to phraseological rather than statistical data.

8.6 Phraseology and grammatical items.

This section ends this chapter on data collection by setting out what methods should be used for the analysis and classification of collocation once collocational patterns have been established. We then attempt to justify the analysis of salient grammatical items rather than the analysis of *any* salient items.

The variety of collocational analysis in literary and linguistic computing leads to a wide range of applications and ways of seeing collocation. For example, Ide (1993) uses semantic images produced by concordancing to study the progression of different scenes in a poem. Miall (1992) uses a similar method to calculate changes in collocations over 10 sections of a single text. However, as mentioned earlier, collocation has been most widely analysed for lexicography, and it is this tradition that this thesis must take into account.

Lexicographers for the first edition of the Cobuild dictionary classify the word into its traditional word class, look at its principal meaning according to context and consider any regular meanings it forms when in close association with other words (Sinclair 1991:72). Word meanings for the dictionary are sorted into broad group of senses, while phrases are also assigned to senses and entries in the dictionary. The general syntax of the word is noted (Clear 1987), for example whether the word is typically followed by a certain semantic or grammatical class of verb, then the field of the word and its stylistics and collocations are noted. This can be seen as a classic approach to collocation: applying traditional grammatical description to newly discovered phraseological patterns that are said to be typical usages of lexical entries. This is a methodology that has been seen to function well for lexicography. But as has been seen, the data very often contradict previous grammatical statements or create new grammatical principles that do not fit the existing paradigm.

Given the emphasis in this thesis on textual and metaphorical reformulation of information, established systemic descriptions of the language would be appropriate to the phraseological patterns that emerge from the corpus. These can be related to patterns found in the reformulation hypothesis (such as complex paraphrase and grammatical metaphor) while more localised processes (such as the salient use of *of* in titles) can be associated with the process of nominalisation and terminology-building. Of course, as stated above, some findings may also challenge the systemic paradigm. The concept of 'phrase' or 'group' for example, may very well be hard to apply to all the patterns we find in the corpus (as seen later). One aim is to establish the broadest systemic principles of the corpus on the basis of the most frequent patterns. But these principles are also related to the writers' inferable rhetorical aims (emphasis of result, evaluation of previous findings and so on) where these appear in the text, and there is thus the possibility of establishing a series of interrelated choices that form a rhetorical system of preferred expression in the texts. Some rhetorical goals are evident from the text and can be related to Swales' rhetorical moves. Others are more subtle and depend on the position of a posture in its textual context (preceding results or following them, etc.).

A final issue that has to be addressed is whether all patterns and words are to be equally valued in the analysis, and whether it is economical to rule out high frequency words as in other collocational studies. It has already been suggested that even if a word is particular to a rhetorical section, then its usage should be explored in the corpus as a whole. On the other hand, this cannot be done for each significant item. One solution may be to take a

sample of high, middle and low frequency items and study their phraseology. However, this would obviate the advantages of the *Wordlist* keyword comparison. It is also unlikely that middle and low frequency items would be equally distributed throughout the corpus and there is strong evidence to suggest that even high frequency items are unevenly distributed across the corpus (as set out in Appendix C).

Previous studies have claimed that high frequency items are stable in use and meaning across the types of language and the assumption is that is if a word is stable it is a 'grammatical item' or a 'function word'. Sager et al. (1980:238) characterise a descending type / token ratio 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 special languages. Phillips also discriminates against 'grammatical items', distinguishing them from "carriers of local meaning in text" (1985:66). There are obvious reasons for this in an automatic analysis of the transparent semantic structure of text, but it is clear that some phraseological patterns involve quite specific semantic content that is distributed among its constituents (as in the case of phrasal verbs such as *set in, set out, set to*) and thus the idea of meaning construction through phraseology and metaphor is obscured and perhaps distorted by low frequency items where more general or even overriding concepts may be omitted from the analysis.

The role of high frequency items in the cohesive system is has been assumed to be minimal. Conjunctions and complex expressions of cohesion are clearly cohesive. But Halliday and Hasan (1976:290) claim 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.". Even given their quite different view of collocation to the one presented in this thesis, when Halliday and Hasan claim that "the higher the frequency of a lexical item... the smaller the part it plays in lexical cohesion in texts" they must also assume that closed-class items are of little interest in the meaning creation of the text since this is at least one of the functions that Halliday and Hasan assign to cohesion.

In contrast, Ljung has argued that closed class items are as revealing as open class items (1991:254). So-called 'closed class' items have been seen to play an important role in expressions and mechanisms of metaphor and reformulation. and it has also been seen (Francis 1993) that the difference between 'grammatical' and 'lexical' items has become blurred, especially from the lexicographers' point of view. While recognising the

differences between high frequency and low frequency items, a conscious decision to analyse grammatical items is undertaken in the data analysis and the reasons for this are set out below. The priority is also to analyse all items equally, especially the phraseology of items that are considered significant to the rhetorical section or corpus in which they are found. This means that at times verbs of different tense or number are analysed separately, and surprisingly they are shown to differ in phraseology. This issue is further discussed in Chapter 11.

PART III: DATA ANALYSIS

In the following three chapters the data obtained for the two specific research hypotheses (Chapter 6) are analysed according to the survey and corpus methodologies set out in Chapters 7 and 8.

CHAPTER NINE: Reformulation in Pharmaceutical Sciences Articles.

9.0 Reformulation in rhetorical sections.

The aim of this thesis is to explain how the author of research articles constructs a new claim in the course of a scientific research article by exploring the development of discourse signalling throughout the various sections of the article. Chapters 9 and 10 attempt to test the reformulation hypothesis:

New cancer research ideas are created by the interaction of two textual processes: 1) grammatical metaphor and 2) discourse signalling (posture).

The survey (Chapter 7) has set out what constitutes an original contribution or claim according to authors, readers and journals in the pharmaceutical sciences. Here, the new claim for each text is discussed using two linguistic elements of reformulation introduced in the literature review (Chapter 4) namely: logogenetic history (progression of grammatical metaphor), and posture (prospection and encapsulation by discourse signals). Logogenetic history is analysed in ten sample texts, its major patterns are set out in summary with a detailed analysis set out later. Posture attempts to characterise the role of discourse signalling in the same texts, and this analysis is set out similarly in Chapter 10. The detailed analysis of both patterns is commented on as the analysis progresses, largely because we find that patterns (especially of grammatical metaphor) are highly idiosyncratic. The question as to whether either process interacts (expressed in the hypothesis above) is discussed in Part IV: Conclusions.

9.1 Logogenetic history.

Logogenetic history refers to the development of grammatical metaphor throughout a text. (c.f. Chapter 4.1, Halliday 1992:70, Halliday and Martin 1993: 217). Halliday defines grammatical metaphor as "transcategorisation..." involving "a new status being conferred" (Halliday and Martin 1993:13) on the deverbalised element. In systemic grammar (Halliday

1985) grammatical metaphor is characterised in terms of two of the metafunctions:

1) ideational metaphor.

This involves the expression of a process as a transitivity role. Halliday (1985:322) gives the example of a mental process (*see*) expressed as complement of a verbal expression (*sight*) in the following:

Congruent expression: Mary *saw* something wonderful.
Metaphorical expression: Mary came upon a wonderful *sight.*

2) interpersonal metaphor

Interpersonal metaphor supplements various aspects associated with the function of the verbal group finite (polarity, modality) and accounts for the complexity of many verbal group complexes in English (expressing modulation: obligation and inclination). While modulation is clearly not a significant factor in cancer research, modality turns out to be very important. Halliday (1985: 336) identifies two congruent and two metaphorical types of modality:

Modulation:	Explicit congruent: I think Mary knows Implicit congruent: Mary'll know
Modality:	Explicit metaphorical: Mary probably knows Implicit metaphorical: It's likely that Mary knows

We see later that many phraseological patterns in the PSC corpus tend towards the implicit metaphorical (or in Halliday's terms 'objective') type of modality. Grammatical metaphor is dependent only on a *potential* verbal paraphrase, which Halliday refers to as the 'congruent form' (1988). He gives examples such as:

#1 *X were <u>developed</u> because they could be <u>used</u> in Y.*

metaphorised as

#2 *The <u>development</u> of X was promoted by their <u>utility</u> in Y.*

Also, grammatical metaphor does not have to be deverbal as in the following example of nominal conversion from an attributive clause and the conversion of quantifier to verb (Halliday and Martin 1993:14):

#1 *X becomes more stable as Z acquires more Y.*
metaphorised as:

#2 *The <u>relative stability</u> of X <u>increases</u> with <u>increasing</u> Y of Z.*

Derewianka (1994) has pointed out that Martin's version of grammatical metaphor includes paraphrase, Halliday and Hasan's (1976) textual reference by general nouns and 'metadiscoursal nominals' (similar to Francis's anaphoric nouns). Derewianka argues that grammatical metaphor should only be considered in cases of transcategorisation of agnate forms where the corresponding transitivity structure can be retrieved. In the analysis below, however, we use Martin's view, especially where metaphors are introduced as ideational themes immediately after a conguent form.

In our text sample, each text typically contains a large number of metaphorical expressions. While it is difficult to single these out, in text CC grammatical metaphor does appear to follow the congruent-to-metaphorical pattern that Halliday suggests. We argue below that this is not always the case. In CC metaphor also appears to progress through a series of systematic steps. In the first instance, the material process of 'decomposition' is expressed as a biochemical (material) process:

...diester 1 rapidly decomposed...

Later in the text, the process is still expressed congruently, but is introduced hypotactically by a 'research' process which approaches Halliday's category of 'mental process':

...,intermediate 4 was shown to decompose...

'Shown' here functions in verb group complexes in the same way that typical mental processes do (for example: 'was *thought/ believed/ seen* to decompose'). Finally, the biochemical process is expressed as a nominal subject of an attributive clause:

The material decomposition of the triester 1 is very similar.

There appear to be two processes at work. Firstly, metaphor takes place in steps rather than in just simple nominal to verbal transcategorisation. Secondly, the global processes that systemic grammar claims are fundamental to transitivity in the grammar are semantically specific: for example, biochemical process (as a type of material process), research process (as a type of mental process). These process types become important in the description of phraseology (Chapter 11).

To summarise, by enabling a process to be reformulated as a participant, grammatical metaphor moves the argumentation from a material process to relational, mental and other discourse- oriented expressions. It is the introduction of new material rather than the

reassignment of transitivity which appears to be a major function of grammatical metaphor. Thus the progression of grammatical metaphor can be seen as a mechanism of change in the textual development of ideas.

9.12 Summary of Logogenetic histories in the text sample.

As mentioned in the chapters on data collection, reformulation is analysed in ten articles (all offered by the authors themselves) (BJ, CC, JCPT[3, 9, 10], CCP, JMC, JNCI, TL, TPS). The sample includes three articles from one author (*MT*'s BJ, JNCI and TPS) and three from one journal (JCPT3,9,10). The sample also comprises the main subgenres represented in the corpus: what the scientists term a 'review article' (a general work in progress report, TPS), three standard IMRD research articles (CCP, JNCI, BJ), four experimental articles with combined results-discussion sections (JCPT3,9,10, JMC) and two communications (CC and TL).

All kinds of progression of grammatical metaphor are possible, although congruent -> metaphorical progression appears to be preferred:

Logogenetic (congruent-to-metaphor): JNCI, TPS, TL, JCPT3, JMC
Cyclic (congruent-to-metaphor-to-congruent): CC, CCP
Reverse (metaphor-to-congruent): BJ, JCPT9
Reverse cyclic (metaphor-to-congruent-to-metaphor): JCPT10

Again we emphasise that these patterns reflect the main reformulations of the central concepts identified in each text: the model does not represent a true picture of all the expressions in the text and can only therefore inform us about the linguistic construction of central concepts. For example: a text may display 'reverse' logogenesis, but this may take place early on in the text, sometimes in the Introduction (in JCPT9, for example), leaving congruent expression for the rest of the text or the reversal may take place in the last sentence of the discussion. Generally, however, we find that changes (one way or the other) occur at the boundaries of rhetorical sections. Here we summarise the major patterns for each text:

1: CC (author *SF*) Bioreversible Protection for the Phospho Group: Chemical Stability and Bioactivation of Di(4-acetoxybenzyl) Methylphosphonate with Carboxyesterase. [Structural Chemistry]

CC displays 'cyclic' logogenesis. The central mechanism of 'release' of a prodrug is

expressed congruently until the results section, then metaphorically in the discussion section. Here it is equated with more specific processes (liberates, trapped) which are then expressed congruently:

Results		Discussion		Discussion
Congruent	->	Metaphor		
released		*release*		
			->	Congruent
				liberated / trapped

2: JCPT9 (authors: *SF,WI,AM,DN*) Bioreversible Protection for the Phospho Group: Bioactivation of the Di(4-acyloxybenzyl) and Mono(4-acyloxybenzyl) Phosphoesters of Methylphosphonate and Phosphonoacetate. [Structural Chemistry]

JCPT9 displays normal logogenesis, although some central biochemical concepts are expressed metaphorically in the Introduction and then congruently thereafter:

Introduction		Rest of text		
Metaphor	->	Delexical	->	Congruent
decomposition		*proceeds with hydrolysis*		*decomposes*

3: JCPT10 (authors *WF,CS,HW*) Latent Inhibitors. Part 7. Inhibition of Dihydro-orotate Dehydrogenase by Spirocyclopropanobarbiturates. [Structural Chemistry].

JCPT10 displays 'reverse cyclic' logogenesis. Here the central concept of 'design' of a new prodrug is at first expressed metaphorically, is then 'unpacked' to allow for new information and is then re-expressed a metaphor.

Introduction		Results		Discussion
Metaphor	->	Congruent	->	Metaphor
design		*design*		*concept of design*

4: TL (authors *JE, JG*) Synthesis of Antiviral Nucleosides from Crotonaldehyde. Part 3.1,2 Total Synthesis of .Didehydrodideoxythymidine (d4T) [Organic Chemistry]

TL displays logogenetic progression from a clinical process to the nominal expression of a research process. The difference from normal grammatical metaphor is that the metaphor is itself metaphorised by lexical reformulation. The central concept of a new methodology is at first expanded (with increasing qualifiers) and then reformulated as a mental (research) process:

Introduction		Results- Discussion
Congruent		Metaphor
can be elaborated	->	*synthesis*
		-> *this route*
		-> *this mechanism*
		-> *this strategy*

[We are assuming, in the above analysis that 'elaborated' is being used technically. The original expression is: *the epoxy alcohol can be elaborated in six steps...*]

5: JCPT3 (authors *WF, CS, HW)* Structural Studies on Bioactive Molecules. Part 17. Crystal Structure of 9-(2'-Phosphonylmethoxyethyl)adenine (PMEA). [Structural Chemistry].

JCPT3 displays logogenetic progression. Central concepts of crystal formation are first expressed congruently and then reformulated lexically as in TL:

Congruent	Metaphorical
(biochemical process)	(general noun)
crystals were formed	*The structure...*
was found to crystallise	*The structure...*

6: JMC (author *PL*) Structural Studies on Tazobactam. [Structural Chemistry]

JMC displays a form of 'selective' logogenetic progression. Certain (infrequent) technical verbs are not expressed metaphorically (flip, puckered, rotate) while others are expressed metaphorically later on in the text:

Congruent	Metaphorical
(biochemical process)	(biochemical process)
hydrolyse	*hydrolysis*
inhibit	*inhibition*
optimize	*optimal*

7: BJ (author *HM, MT*) Metabolic substrate utilization by tumour and host tissues in cancer cachexia. [Cancer Histopathology]

BJ displays 'reverse' logogenesis. Sometimes the change is in the abstract (as in *utilisation*) at other times the discussion section (*oxidation*). In the discussion section, there is also reverse cyclic expression where *utilisation* is reformulated as *consumption* but congruently reexpressed as *utilized*:

Congruent	Metaphor
oxidation	*oxidises*
utilization	*utilized*
consumption	*consumed*

8: JNCI (author *MT*) Lipolytic Factors Associated With Murine and Human Cancer Cachexia [Cancer Histopathology].

JNCI displays typical logogenetic progression.

Congruent	Metaphor
inducing /induces	*induction*
glycerol released	*glycerol release*
associated with LMF...	*cachexia-related LMF*

9: TPS (author *MT*) Newly identified factors that alter host metabolism in cancer cachexia [Cancer Histopathology]

TPS displays typical logogenetic progression, as with JNCI with a possible example of modularity:

Congruent	Metaphor
cachexia could arise...	*the possible interrelationship*

10: CCP (authors *YW*) Relationship between the melanin content of a human melanoma cell line and its radiosensitivity and uptake of pimonidazole. [Cancer Histopathology]

CCP displays 'cyclic' logogenesis. As with other 'stepped' types of metaphorical progression, the steps can also be enabled by lexical reformulation rather than just transcategorisation:

Congruent		Metaphor		Congruent		Metaphor
accumulate	->	*uptake*				
	->	*concentration*				
			->	*bases can concentrate*		
					->	*depends on concentration*

Here, metaphor at first reformulates accumulation as *uptake* which is itself reformulated to *concentration* (as a measure of uptake). This is then rendered congruently and then reformulated nominally.

9.13 Preliminary conclusions on grammatical metaphor.

The following preliminary conclusions can be proposed:

1) Logogenesis does 'progress' from congruent to metaphorical in texts. However it can progress in cycles or even in the reverse direction. This suggests that argumentation in some research articles is less linear than explanatory basic scientific text.

2) Logogenesis reformulates technical and biochemical processes most of all, the majority of empirical and research processes being expressed congruently. We suggest that continuing congruent expression or reversal of logogenesis signals that a biochemical process is negotiated at that point in the text or is 'at stake'. On the other hand, metaphorised expressions are incorporated into the accepted paradigm: they are 'assumed'.

3) With a small sample it is hard to say whether direction of logogenesis corresponds to journal, genre or any other variable, such as author or topic. On the basis of the above limited results, there is no correspondence. This would suggest that logogenesis is linked to global *'genre'* (Martin's 1993 sense) such as 'explanation' or 'report' rather than to specific moves. This hypothesis is further tested below.

We have oversimplified the results to emphasise the patterns of logogenesis. In order to discuss the nature of the 'claim' or 'novel science' in any of these texts it is necessary to go into more detail. We have already found that grammatical metaphor often involves complex relations between semantically related sets of lexical chains in these texts, and any attempt to link reformulation with progressive steps in scientific argumentation necessarily involves a closer discussion of the text. In the following section, detailed analysis of grammatical metaphor in each text above is taken up, especially to assess the role of reformulation in the expression of scientific claims.

9.21 Cyclic Logogenetic History in text CC.

CC is a communication (1520 tokens) with a prototypical IMRD rhetorical structure. In CC the expression of the central concept, the breaking up of the compound (methylphosphonate), is referred to as *chemical stability* and *bioactivation* in the title. The essential novelty of the article depends on the observed release of an electrically charged

compound in a series of reactions with esterase, and this central concept is expressed in the Discussion section by a complex nominal group as subject of a material research process (outlined) in §50: *The potential problems associated with the release of a highly reactive benzyl carbonium ion have been outlined....* This expression is constructed in a number of parallel steps within the text, and we refer to it here as the 'target phrase'.

In the first step, *release* is at first expressed congruently as a material process (§27) *to give either the monoester... together with the...carbonium ion.* The carbonium ion is then implicated as a key entity in the activation of a prodrug for the first time in the Results section (§27) and it is later premodified as *benzyl carbonium ion* in the Discussion section (§41). The next step is to establish the reactivity of the *carbonium ion*, which happens in §34, the first sentence of the Discussion. Prior to the Discussion section the reactivity of the carbonium ion is expressed by a congruent material process (*proceeds, will assist in, decomposes*) or by a semi-metaphorical delexical verb (*undergoes*: a very frequent expression in the corpus):

...will assist in cleavage, (congruent)
...the reaction proceeds via the benzyl carbonium ion with C-O cleavage. (congruent)
...rapidly decomposed in less than 3 min. (congruent)
...shown to decompose... with a half-life of 17 min. (congruent)
...undergoes 88% solvoyosis. (semi-congruent / delexical)

However, in the Discussion section, the high reactivity of the carbonium ion is converted into grammatical metaphor: *The ready removal... of the 4-acetoxybenzyl groups with carboxyesterase* (echoing a phrase in §27), where *carboxyesterase* has already been established as decomposing *via* a carbonium ion. Thus the high reactivity of the carbonium ion is formulated congruently in the Results section, and metaphorically in the Discussion section until the target phrase, where we have the qualifier *of a highly reactive carbonium ion.* Clearly, the steps are not transparent: it is necessary to link this metaphorical progression with a fact established elsewhere in the text: the carbonium ion decomposes *via* carboxyesterase.

In the third step towards the target phrase, *release* is equated in a parallel expression with the congruent material expression of removal in §9: *although some...phosphate is released, the second...group is removed only slowly.* The fourth step takes account of the fact that *release* has already been formulated nominally in expressions qualifying a problem: *problems of formaldehyde release* (§11). Towards the end the expression include modality: *could ... provide a sustained release of parent drug* (§35). The modality itself is

nominalised in the final nominal: *potential problems associated with the release of a highly reactive benzyl carbonium ion* ($50). Subsequently, the Results discussion establishes the reactivity of the carbonium ion, while the Discussion section equates the location of the carbonium ion being congruently *trapped* ($36,39,45,47,50) with *liberated* ($37=*released*). However, the expression in $50 is the first point at which *release* refers to exchange of ions, where previously chemical process verbs refer to whole groups or compounds (*decomposes to X ester, ester y is removed by hydrolysis, removal of group Z*). By $50, then, CC has moved specifically from the discussion of *carboxyesterase* (and other complex compounds) to a specific quality of the compound: its characteristic ion loss (or *release*). In addition, *release* can be seen not only as a nominalisation of previous verbal expressions of chemical processes, but as an instance of a word which simultaneously infers complex processes of chemical exchange (*trapping, release, liberation,* and *decomposition*). It is possible that pharmacologists can make these connections directly, but it is our contention that these very small connections are established solely by the text.

Finally, we note that the complex nominal group of our target phrase is subject of a research verb (*have been outlined*). Referring to difficulties with highly reactive compounds as *problems* allows the author to refocus her (and in $11 other') observations and relate them to research processes, usually reporting verbs (*led us to consider, have been outlined*). As we have noted, this also functions as interpersonal grammatical metaphor, allowing for the expression of non-relational processes (verbal, material etc.) in order to enact reformulation.

9.22 Logogenetic history in text JCPT9.

Relational processes are important in explanatory rhetorical moves in JCPT9. The nominal expression of material processes in text JCPT9 can be seen in expressions involving *hydrolysis, decomposition, degradation* and *cleavage*. As with Halliday's model, all three processes are expressed verbally when first mentioned:

$3 *...both series of compounds hydrolyse with half lives of....*
$8*...the...ester...was not cleaved by PLCE.*
$24 *Although the diesters do degrade further to give benzyl and phenyl phosphate...*
$43 *The monoester decomposed further to...*

Cleavage is expressed as a nominal for the rest of the article:
$11 *...facilitates the cleavage of the B-O bond....*
$18*...very resistant to cleavage...*
$20*...subsequent P-O bond cleavage...*

Decomposition is expressed once again nominally and once verbally. Similarly, *degradation* is expressed nominally throughout the RD section in connection with processes that express delexical reactions (\$48) or with processes that express the empirical circumstances of the reaction:

\$48 *...their degradation must first proceed via hydrolysis...* \$83... *first order degradation, when the degradation rate constant is...* \$97 *This degradation follows a pathway...* \$98 *...data collected during the degradation of...*

Congruent expression of the reaction takes over again towards the end of the RD section:

\$91 *...which then spontaneously degrade to the diesters...*
\$122 *The diesters were found to degrade to methoxycarbonylmethylphosphonate...*

This verbalisation is mirrored in the last two cycles of the RD section by phrases that were seen to be semantically related in text CC, in that they are obligatory sub-processes of *decompose*, and are themselves closely related:

\$126 *Studies are underway to replace the methoxy group with a substituent that could be removed...*

\$130 *with the phospho moiety being liberated.*

In most cases, verbal expression of material process equates with congruent expression. However, as we have seen above, there are cases where we can argue for an intermediate stage of logogenesis. In the abstract *Hydrolysis* is 'deverbalised' where it is expressed nominally without an article and with a delexical verb 'undergo':

\$4 *the monoesters undergo chemical hydrolysis...*

Even though the normal congruent expression would involve a possible ergative verb '*monoesters hydrolyse*', this usage still appears to be congruent, where the process is realised by a delexical verb. The use of a more semantically neutral verb '*proceed + with / via hydrolysis*' may also be considered delexical where the path of hydrolysis is more important than the hydrolysis itself:

\$10 *...both the chemical and enzymatic hydrolyses... and the PLCE hydrolyses of esters proceed* via *hydrolysis of the acyl group to give the acylate anion...*
\$48 *their degradation must first proceed with hydrolysis...*
\$91 the *PLCE-catalysed hydrolyses* proceed via the 4-hydoxybenzyl intermediates

We can consider the forms with 'undergo' and 'proceed' as intermediate because they can be contrasted with other nominal expressions of *hydrolysis*, which allow the formation of complex nominals, with either a left-branching head (*rate of*) or modifier (*chemical*):

\$7 *rate of enzymatic hydrolysis was most rapid for...*
\$10 *chemical and enzymatic hydrolysis of (...) esters and the PLCE-catalysed hydrolyses of the di(...)esters* [...proceed via hydrolysis...]
\$23 *The rate of enzymatic hydrolysis can be controlled... derivatives undergoing only slow hydrolysis.*
\$30 *...phosphonoformate was highly reactive towards chemical hydrolysis...*

Mentions of hydrolysis after \$61 (*the monoanion... is reported to hydrolyse with P-O cleavage*) are verbal, and reflect the general topic shift away from the entire process of synthesis to reported observations of the charged compounds in the reaction. The tendency to introduce new lexical verbs or re-express verbally processes such as 'degradation' in the later RD section may be characteristic of a more typical 'Discussion' section.

9.23 Reverse cyclic logogenetic history in JCPT10.

JCPT10's contribution to medical science lies in the *design* of new compound prodrugs (inhibitors) to aid chemotherapy. While phrasing of 'design' is largely metaphorical in the text, it is unpacked and rephrased during the text and new elements are associated with it which then appear to form a new phraseology. While JCPT10 does contain typical cases of simple transcategorisation (\$24 *were all bound* \$26 *The tightest binding*) we see that nominal expressions of material processes (design) are themselves involved in considerable reformulation.

The concept of a design of potential structure can be inferred from a complex nominal from the Abstract (substrate surrogate = template = design) :

\$5 *the possibility of using substrate surrogates as templates for constructing latent inhibitors.*

Throughout the majority of the text, *design* is expressed metaphorically, while the compound itself comes pre-packed from a previous reformulation in \$7 (*inhibitors of a wide variety of enzymes*):
\$8 *The goal is to establish design methodologies that will make it possible to obtain highly selective enzyme-activated inhibitors...*

Later, *substrate surrogates* are associated as *inhibitors*

$10 *substrate surrogates... might provide a general design concept for highly selective enzyme inhibitors.*

At a key point in the text, 'design' is unpacked as an embedded verbal process where the compound is associated as an inhibitor:

$12 *compounds designed as potential inhibitors...*

In the Design / Methodology section *surrogate* is incorporated as a premodifier as *the surrogate substrate concept.* It is only towards the end of the Results-Discussion section that the design concept is reformulated congruently:
$52 *substrate surrogates can be designed*

This is in turn nominalised as $53 *demands inherent in the design concept*

and finally the design concept is reformulated more broadly as a *strategy*:

$54 *distancing from the natural substrate is likely to be the best strategy in design, a concept that we are pursuing...*

Reformulation within nominal processes involve the incorporation of new elements as postmodifying clauses or phrases and then as premodifying qualifiers. This variable expression of a concept, and the apparent progression from general to particular (and then outwards again to the broader 'strategy') appear to confirm Halliday's suggestion that grammatical metaphor (aside from translating process as participant) allows for tighter nominal 'packing' and that this process is incremental throughout a text. In texts JCPT3 and TL, also, we see a process of gradual nominal packing followed by nominal unpacking in the Results-Discussion section.

9.24 Logogenetic history in JCPT3.

JCPT3 belongs to new field of applied molecular mechanics, where the aim is to provide accurate chemical structure description of a drug in order to evaluate its pharmaceutical properties elsewhere. This description, as well as the medical application, is encapsulated by the nominal *The crystal structure of the antiviral agent* in the Abstract $2. JCPT3 gives few clearcut examples of transcategorisation as grammatical metaphor, although there is much global nominal reformulation. Stuctural measurements reported in $24 are rephrased in terms of metaphorical modality: $26 *the present structure provides a concrete representation of the likely intermediate.* Other indirect metaphors involve the metaphorisation of a process expressed as a (semi)technical verb (of which there are many in this article) to a superordinate term such as *structure*, as in the following paris:

$10a *Crystals ... were formed,*
$10b *and it is the crystal structure of PMEA that is discussed*

$15 *Reflections were collected to 2qmax = 50o, yielding 2433 independent reflections.*
$16 *The structure was solved by direct methods...*

$21 *PMEA was found to crystallise as the zwitterion protonated at N1.*
$22 *The most unusual feature of this structure...*

$24 *The close 1,4 contact that arises between C8 and C11 is relieved by expanding angle C(10)... to a value 7.4*
$25 *.. the present structure provides a concrete representation of the likely intermediate.*

Derewianka (1994) and Martin (1993) accept this as a form of grammatical metaphor: in fact they see this as more significant than transcategorisation involving the same lexical item. It is certainly referential - Halliday and Hasan (1976) would account for this as a case of 'collocation' (general or related noun reference). Again, this type of topical reformulation becomes particularly salient as evidence of encapsulation and lexical reformulation (Chapter 10).

9.25 Logogenetic history in JMC.

JMC belongs to a similar field to JCPT3 (although by different authors), and can be seen to use a similarly diverse set of technical verbs. JMC presents some interesting cases of technical verbs which are not nominalised directly in the rest of the text. One of these is $17 *the thiazolidine rings are puckered...* $32 *The thiazolidine ring was less puckered...* The other verb *flip* first appears at the end of the Abstract:

$7 *Molecular mechanics supports the hypothesis that the carboxyl group can rotate freely and the triazole cap 'flip'.*

and again in the second Results-Discussion section:
$39 *In the case of the primed molecule slightly more energy is required to 'flip' the triazole ring through 180°.*

Puckered appears to be incorporated into the subsequent text by 'conformation' and finally by a general encapsulation *this geometry* ($21) and so is indirectly metaphorised. *Flip* is introduced as a pseudo-term, and can be seen to be nominalised as *rotation* in $40 although later this is reverbalised as a postmodifying participle *rotated through* in $43 and in $46. This suggests that some processes remain largely congruent throughout the text, and these

can be contrasted with those in JCPT9 above (*hydrolyse - hydrolysis*) where the process is nominalised in the text after first mention in order to allow for the incorporation of new information. In JMC this includes *inhibition* and *optimization*: \$10 *have been found to inhibit many lactamases...* \$11 *wide range of inhibition and weak induction of lactamase..* \$13 *this lactamase inhibitor...* \$29 *The optimum geometries of both molecules were determined...* \$30 *The independent structures optimized to .. identical conformations* \$32 *was less puckered after optimization.* It may be that, just as some nominal items in lexical chains are more 'at stake' than others, so processes that remain expressed as verbal processes (*flip, puckered, rotate*) are 'at stake', while others (*hydrolyse, inhibit, optimize*) are 'assumed' - especially after first mention, thereafter taking on the functions of grammatical metaphor observed above.

9.26 Logogenetic history in TL.

TL's claim is based on the creation of a new compound, referred to in the title and Abstract as *total synthesis of the antiviral agent d4T (from crotonaldehyde).* It can be seen that the *synthesis .. from crotonaldehyde* becomes 'unpacked' during the course of the introduction section. \$7/\$8 introduce postmodifying adjuncts of circumstance: *syntheses from nucleoside starting materials... non-chiral pool materials.* \$8 *syntheses of the anti-AIDS drug AZT ... from the inexpensive achiral starting material, crotonaldehyde.* In the results-discussion section *d4T* is replaced from the original formulation by qualifiers indicating its function and derivation instead, and the head noun and modifier are further premodified, this time with evaluative epithets: \$22 *the efficient synthesis of a range of important antiviral modified nucleosides from cheap achiral starting materials.*

This kind of additive nominalisation is consistent with the general patterns observed in the other texts; however TL also displays another pattern of grammatical metaphor which is less clearcut but fundamental for the analysis of reformulation and encapsulation. Throughout the text, the key process of synthesis is lexically reformulated to express a different stage of synthesis, with a general elaboration towards the Methodology section and general reconstruction towards the Results-Discussion:

\$7 *A number of syntheses...*
\$8 *... novel and versatile synthetic routes...*
\$9 *... epoxy alcohol can also be elaborated in six steps...*

The six steps are then presented as a cycle of very specific nominalised material processes:

\$11 *Ring opening of the epoxy alcohol...*
\$13 *Cyclization .. proceeded in near quantitative yield...*
\$14 *Combination of the glycosides obtained from this reaction... during the ring-opening reaction...*
\$15 *Acetylation of these alcohols...*
\$17 *Treatment of the seleno compound*
\$18 *Deacetylation was effected...*

The process is then re-generalised:

\$20 *This route provides d4T in six steps...*

And in \$22, the process is reformulated first as a *methodology* and then as a *strategy*, concepts with a much broader scope than *synthesis:*

\$22 *The completion of this total synthesis... establishes this methodology as a general and versatile strategy towards the efficient synthesis of a range of important antiviral nucelosides...* \$23 *Further work on the extension of this methodology... is under way.*

Text TL provides us with a very clear pattern of reformulation: the methodology is nominally packed, congruently unpacked and at the same time it is reformulated in terms of increasing claims from *synthesis (*via *steps* and *route)* to *methodology* and finally to *strategy*. This is the same process observed in JCPT10. While the first three items are types of topical reformulation, the last two are anaphoric nouns refocussing the methodology as a higher order of research activity. These are, again, at the boundary between grammatical metaphor and key discourse items in the structure of the text.

9.27 Reverse logogenetic history in BJ.

In BJ, the *utilization* of sugars by the brain is the object of study, with the claim that this is higher in animals with cancer. In the Abstract, *utilization* is at first nominal then verbal in the final sentence (*\$10 ketone bodies may be utilized as a metabolic fuel*). In the Results section, the gradual building up of information into the noun group (research quantifier *decreased/increased* and classifier *brain*) can be seen in the following examples of nominal metaphor:

\$63 *Glucose utilization ... was significantly greater than...*
\$64 *There was no difference in glucose utilisation by the MAC 16 tumour...*
\$70 *was accompanied by a marked decrease in glucose utilization by the brain*
\$81 *The decrease in glucose utilization by organs in tumour bearing mice...*
\$89 *This [...] was accompanied by a marked increase in both lactate and hydroxybutyrate utilization in the brains of animals bearing the MAC16 tumour...*

\$70-\$89 mark a high point in the nominal packing where *decrease* and *increase* act as head

of the nominal group, while by \$90, \$93 and the Discussion section: \$102, 103 these have been integrated as an epithet with *utilization* as head:

\$90 *This increased 3-hydroxybutyrate utilization in the brains of tumour-bearing animals is probably ...* \$93 *an increased utilization of 3-hydroxybutyrate*, \$102 *decreased glucose uptake* \$103 *decreased glucose utilization.*

In the Discussion section, *utilization* is explained in terms of a synonym: *consumption* (\$96) and associated with new processes (such as*suppress*) (\$105,119). It is finally rephrased verbally (\$120):

\$99 *consumption ... is suppressed by pleural effusions...*
\$105 *The inability of glucose loading to suppress oxidation of fatty acids... suggests that...*
\$108 *In addition, if some glucose is not oxidized directly but is first converted into fat before supplying utilizable energy...*
\$120 *...lactate has also been shown to be utilized by a number of rat tumours...*

This may indicate that as a major topic change in the text takes place, grammatical metaphor is unpacked (or reverbalised) to allow for new formulations to be incorporated. Here, as in topic changes in CC and JCPT9, the topic becomes more specific, the new view (in the text) of *consumption* as *oxidation* being a significant change in perspective. At the same point in the text, verbal expressions of new processes are accompanied by projection (infinitive and participle clauses embedded in nominal groups) such as in \$105 and \$120. In \$125 this is finally reformulated lexically as *increased metabolic requirements*:

\$125 *Thus cachexia in the host could arise from an inability of the host to adapt to the increased metabolic requirements of the tumour-bearing state.*

In BJ, the first of our cancer research articles, the claim realising *oxidation* as a characteristic of the tumour is made by the interaction of grammatical metaphors, allowing for projection and, as we have seen above, formulating the participants' new relationship with non-material processes (verbal process; reporting). It has already been observed that projection is a grammatical process that is characteristic of the 'explaining genre' (Martin 1991). As in other Discussion sections, the rhetorical process of explanation may also coincide with a 'reverbalisation' of the text. These processes indicate that grammatical metaphor varies across the text -not just in a linear verbal - nominal transformation, but also as a function of the rhetorical needs of the text. Halliday's example texts are in fact largely explanatory, in other words they do not change 'genre' (in Martin's sense). A linear progression from congruent expression to grammatical metaphor may be a property of

single 'genre' texts, while the texts we deal with change from report to explanation in several cycles. In the case of BJ this hypothesis can be further tested, since *MT* deals with the same subject in the highly 'explanatory' text TPS (9.29 below).

9.28 Logogenetic history in text JNCI

JNCI covers a similar area to BJ (involving the same author), except that here the finding is that cancer cachexia is associated with a lipid factor, a molecule involved in the transfer of information at the cell membrane. The factor is *induced* by cancer (\$4), while the tumour *releases* a measurable lipolytic factor (\$17). Again, these are expressed at first as key congruent processes in the argumentation of the text, and follow a typical verbal-to-nominal pattern: \$4 *inducing* \$ 15 / \$17 *induces* \$75 *induction* ... Also, \$20, 30, 32 are all instances of linear progression of metaphor: *glycerol released* \$64 *to measure glycerol release.*

However, the association between cachexia and cancer is not a material process that can be traced as simply as the ones mentioned above, and as it is expressed in text JNCI does not follow a straightforward congruent-to-metaphorical pattern. The relation is introduced intertextually from previous research, but expressed congruently at first in the material process *mediates* (\$8). The expression of the cancer / factor relationship is then expressed in attributive relational clauses, where the attribute *associated* followed by a preposition encodes the a relationship as a process but does not signal causality:

\$17 *Weight loss is associated with the presence of a catabolic factor*
\$10 *Weight loss is accompanied by marked anorexia*
\$19 *lipolytic factor associated with weight loss.*

The same relationship is expressed metaphorically to allow for a reporting process verb to encode the relationship in \$12: *TNFa showed no correlation with weight loss.* However, the expression of association disappears in the Methods and Results sections, only to reappear as attribution again in \$71 *The lipolytic factor described in this report is closely related to the cachectic state...* By \$52 in the Results-Discussion section the relationship is expressed as a complex classifier in the premodified nominal *the cachexia-related lipolytic factor.* In the latter part of this section, the expression is reverbalised or expressed in attributive relational clauses:

\$58 *Cancer cachexia is characterized by a marked depletion of lipid stores...*
\$65 *The activity of this material differs from the tumor lipolytic activity...*

\$69 *The cachexia-associated lipolytic factor also differs...*
\$71 *The lipolytic activity described in this report is closely related to the cachectic state...*

In \$76 the association is formulated as grammatical metaphor in a complex nominal within a verbal process clause:

\$76 *These data ...suggest a major role for the tumour lipolytic activity in the development of cachexia.*

In JNCI we have evidence that empirical or research processes (often expressed by relational clauses) are metaphorised just as material processes are, and that the logogenetic direction again progresses from metaphorical (in the reporting genre in earlier sections of the text) to congruent (in the explaining genre of the latter Results-Discussion section).

9.29 Logogenetic history in text TPS

TPS is a review paper, reporting the findings set out in JNCI and BJ to a wider audience. We have seen above that while the text by the same author (BJ) displays a heterogenous reverse (congruent-metaphorical) pattern, text TPS has a linear pattern. In previous articles the researchers observed a molecule produced by the tumour (expressed in \$7 here). TPS attempts to move the argument outwards to the relationship between *cachexia* and the tumour's effect on the metabolism. Thus the claim progresses through three cycles, increasing in specificity and strength within each:

1 a) Cachexia is a process *associated with* cancer (growth) (\$9,14)
b) Cachexia is *essential* for tumour growth (\$19)

The search for LMF is initiated in \$21 *One possible mechanism for this stimulation could be...LMF:*

2 a) LMF is involved with cancer (\$21 phrased as *stimulation of*)
b) LMF causes cachexia (\$28, 34, *is responsible for / is involved with*)

EPA is introduced because it may act as an inhibitor of LMF (hypothesised in \$34):

3 EPA reduces tumours (\$42b, 43: *EPA displays antitumor activity... possible to produce antitumor drugs*)

The initial claim is first expressed congruently (although markedly hedged by modals) in the abstract: \$6 *This suggests cancer cachexia could arise from the metabolic effect of the tumour...* and then reformulated metaphorically (we also note the metaphorical

reformulation of the modality...) in §8: as *Figure 1 shows the possible interrelationship between tumour and host metabolism that could account for the development of cachexia.* The relation between *starvation / nutrition* and *increase in body weight / progressive weight loss* is made by 'empirical' process verbs, where 'empirical' indicates items that are used to express cause and effect or observed qualities of chemical reactions. These verbs are neither frequent in the text as a whole, nor nominalised as grammatical metaphors later on, and include from text TPS: *leading to some increases, could arise from, could account for, lead to breakdown.* Those verbs in the abstract that are metaphorised in the main body of the text are material, specifically biochemical, processes (from the abstract: *attempts to reverse the wasting process, a peptide... which is synthesized and released, catabolic factors produced by the tumour*). It would seem that our hypothesis is correct: the informative, explanatory nature of TPS appears to correspond with a homogenous congruent-to-metaphorical progression.

9.210 Cyclic logogenetic history in text CCP.

In CCP it is claimed that a drug used in conjunction with radiotherapy (PIMO) is not suitable for the treatment of skin cancer. This finding relies on the fact that PIMO did not *accumulate* sufficiently in the target cells, a process also nominalised as *uptake*, signifying that cells actively take in PIMO. In the abstract this is expressed at first nominally and then congruently with the epithet reformulated as a premodified prepositional phrase at the end:

§2 *The intra-cellular uptake of the weakly basic radiosensitiser pimoniadazole...*
§7 *This increase in the cellular uptake of PIMO....*
§11 *In conclusion, PIMO accumulates in very heavily pigmented melanoma cells...*

At first expressed congruently in the introduction (§16) *accumulates* is then re-expressed as *uptake* (§23,24) and in the Methods section as *tumour uptake, cellular uptake,* (§31,39,42). While metaphorically expressed as *uptake* the process corresponds with the congruent (verbal / relational) expression of quantification, as in *increase tumour uptake, cellular uptake was measured/ analysed.* In the Results section, the process as metaphor has acquired a series of specifications: *uptake of PIMO into hypotonic phase Na+ cells.. was compared* (§68). By the Discussion section, the transitive process of *uptake* is de-emphasised and re-expressed as *concentration* and also *content*, while PIMO is generalised by molecules that share its properties: *weak bases can concentrate in melanin- containing cells* (§129) and then nominally: *intracellular concentrations of weak bases* (§130,131,132,134). The process varies between the original verbal expression of the

Introduction: *PIMO accumulates in tumours/ tumour cells* ($136,139,144) and nominal expression: *accumulation (of PIMO)* ($139,140,143,134, 144,146,147). In the Discussion section of the text ($130-147) the claim is explicitly restated with strong evaluation: $132: *the results clearly show that... concentration of PIMO depends strongly on intracellular concentration...* There appears to be a pattern of logogenesis here, but it is more akin to the heterogenous use of metaphor in some of the structural chemistry texts. The observation of grammatical metaphor has also been obscured by the fact that reformulation of terms is also happening at the same time: the move from 'accumulation' to the expression 'concentration' in this text demonstrates the complex series of connections that have to be taken into account even when a single idea is being followed throughout the text. We can claim, as in text BJ, that the processes of reverbalisation and variation of grammatical metaphor correspond with the rhetorical construction of the scientific claim, but changes in expression do not appear to correspond with identifiable rhetorical moves.

9.3 Summary of logogenetic history.

We have seen that in all the sample texts grammatical metaphor contributes to reformulation of the topical elements of the text and corresponds with other grammatical systems (clause relations and modality, among others). The mechanisms of reformulation can be seen in the following observations about the text sample analysed so far:

1 Grammatical metaphor increases in the text, but generally only involves material (biochemical) processes that are textually 'assumed'. 'At stake' processes remain typically congruent, or in variable distribution in the text. The expression of empirical processes is more likely to remain congruent.

2 Grammatical metaphor allows for the introduction of new information:

- an increasing specification of the nominal, non-congruent expression of biochemical processes (nominal 'packing').
- the expression of (material) process as participant allowing for reformulation in terms of the authors' own research (empirical and research processes typically expressed by verbal, cognitive and relational processes).

3 The progression of grammatical metaphor appears to vary with the varying status of the claim in the text. A change in explaining or reporting sections and subsections corresponds with congruent expression of a previous metaphorical formulation. The change also corresponds with a refocussing of previously 'packed' items, either 'unpacking' them or assigning them new relations.

CHAPTER TEN: Posture and Discourse Signalling.

10.1 Posture.

The second hypothesis postulates that reformulation plays a key role in the construction of new science. This chapter describes the variable use of discourse signalling on the same sample of research articles analysed in Chapter 9 for grammatical metaphor. The link between metaphor and posture is not formalised until Part IV, although similarities emerge and are commented on below.

In the posture model, the sentence is seen as the state of the discourse supported by mechanisms of maintenance or change (c.f. Chapter 4.3 , Sinclair 1980, 1993d). We have seen that whereas cohesion establishes the principle of links with specific referents, Sinclair has argued for an alternative view of cohesive relations where "each new sentence makes reference to the previous one, and encapsulates the previous sentence in an act of reference" (1993d:8). Posture includes many cohesive categories that are the same as those established by Halliday and Hasan (1976, 1989). However, we identify three major particularities that the posture framework does *not* share with cohesion:

1) Posture functions across clause complexes (coordination and subordination).
2) Posture gives more prominence to reformulation than to conjunction.
3) Certain postures have precedence over others.

As noted earlier, the posture model also unites Francis's (1985) model of anaphoric nouns and Tadros's categories of prediction. Sinclair reassigns these, together with Halliday and Hasan's cohesive relations, into four main types with several subtypes (those we use in the analysis below are underlined):

1 Encapsulation.

A: **Logical acts**. The sentence encapsulates a previous one, either <u>implicitly</u> (the default for all non-initial sentences where a 'logical' relationship can be inferred) or <u>explicitly</u>. Implicit acts are 'logical' because the onus is on the reader to infer coherent relations. The implication is that if there is no cohesion, one can assume that there is still some relation acting in the discourse and that the reader (or hearer) is expected to build these relations

more autonomously than with help from explicit signals (hence the relation between explicit signalling and the interactive rather than the autonomous plane c.f. Sinclair 1981). Explicit signals occur most typically with conjunctive discourse items such as *yet, also, therefore, so, in fact, as a result, consequently...* (this includes Winter's (1977) Vocabularies 1, 2 and 3). Ellipsis also counts as explicit logical encapsulation. Explicit encapsulation may also involve 'internal' encapsulation of a previous clause, rather than a previous sentence.

B: **Deictic acts**. A specific element in a sentence encapsulates a previous sentence or part of a previous sentence by either rephrasing or refocussing. Rephrasing involves encapsulation of entire propositions by superordinate lexical items (*things, these data... these compounds*) or items which label propositions as illocutionary / non-illocutionary utterances, cognitive processes, types of text or evaluative 'facts' (*this request, this example, this opinion, these findings, this problem*). These correspond to Francis's (1985) anaphoric nouns. Sinclair uses the term 'deictic lexical' encapsulation for all of these. Refocussing differs from rephrasing in that it involves an overt relationship between the lexical item and the previous discourse. The main instances of refocussing involve 'selective' deictic acts, consisting usually of pronominal / demonstrative reference to specific items in the immediately previous discourse, or simple / complex repetition e.g (*this (+ repeated item), one of these, such a...*). Including deictic acts comprise demonstrative statements which reformulate previously stated propositions. As we interpret it, inclusive encapsulation appears to be confined to the use of *this* as head rather than modifier.

2 **Prospection.**

Prospection occurs "where the phrasing of a sentence leads the addressee to expect something specific in the next sentence." (Sinclair 1993d:12). Prospection includes attribution (*not* the same as 'attributive clause') where quoted speech or propositions are introduced by a verbal process (*the statement that..., his message... is reported* and typically in our sample: *Numerous studies..., In this study...*). We also find that projecting clauses can be involved in prospection, as in Tadros's (1985) 'reporting' category of prediction.

The second type of prospection is advance signalling (akin to Tadros's categories of advance labelling and enumeration) where a proposition is given an abstract label (as opposed to a verbal label) and can be expected to be expanded in a further sentence (*There are two reasons for this... The implications are daunting...*). Topic selection occurs where

the writer introduces a new argument to the text, and where the choice of new subject matter is certain to be taken up at a later point. This appears to include Tadros' (1985) categories of recapitulation, hypothetical prediction and rhetorical question. The prospected sentence is labelled 'prospected' with no further posture function.

In our text sample, we have marked topic-oriented titles as prospective topic selection (e.g. *Design and Synthesis of Inhibitors, Inhibition of Dihidro-orotate dehyrogenase* etc.) and this makes the initial sentence of the sub-section 'prospected'. Text-oriented titles (*Introduction, Methods* etc.) have not been assigned posture labels although they presumably do prospect as some kind of 'advance signalling'. First sentences introduced by text-oriented titles have usually been labelled 'text-initial' and are mostly non-prospective. For example (from JCPT9):

Introduction
(No Label)

Drugs that are charged at physiological pH often have limted cellular penetration which necessitates large intravenous doses to achieve a therapeutic effect.
(TI - Text initial)

Some text-initial sentences do prospect, at least internally, as in text CC:

Discussion
(No Label)

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.
(Prospection, involving verbal echo back to the main text)

Hoey (personal communication) has pointed out that it this kind of formulation, involving grammatical projection, may be highly genre-specific. Other assignments within the model may reveal areas of posture that are also very specific to the genre and

3 Verbal echo.

Verbal echo involves distinct lexical repetition, where the discourse function of the sentence may be either prospective or retrospective. This is an exception to the model, although it could be argued that it allows for long-range cohesion to be used in the formulation of posture. It appears to be common in political discourse in, for example, Majors's 1995 speech to the Conservative Party conference: *$1 The best route to jobs is more small businesses. $2 We are the party of small businesses.* The message is thus more important than the act of specific reference where the change in theme introduces a new posture.

4 Overlay.

Overlay involves paraphrase and a grammatical parallel between two sentences where the second is either more or less focused. Overlay is always considered to be encapsulation. Sinclair includes this 'exception' because selective deictic acts and verbal echoes do not deal with paraphrases that appear to 'colligate', that is echo grammatical structures. Sinclair's example includes *...by studying [the products] of their rivals...* which is then rephrased more generally: *so that they can learn to keep in touch with trends in other countries.* From John Major's speech we find: *$1 Where were they when we cut inflation? ...$7 Like McCavity, Labour wasn't there.*

10.2 Posture in the Appendix. Below is a summary of the prospection and encapsulation patterns as they have been described above. They are marked next to each sentence of each sample text in Appendix B (parts 1-10). To facilitate reading, each instance of explicit encapsulation is underlined, while lexical reformulations are in **bold**.

TI -Text initial [No posture]
E - Encapsulation (the default is 'logical implicit' and is not marked)

x - Explicit
e - Ellipsis

d - Deictic, including one of the following:
ls - Lexical refocussing (also known as 'selective')
lr - Lexical rephrasing
i - Including

P - Prospection
ts - Topic selection.
at - Attribution.
al -Advance labelling.
p- (Sentence prospected by one of the three previous categories).

VE - Verbal Echo

O - Overlay

Note that in clause complexes, the first clause is referred to as a) and the second as b) and that if subordinate clauses have different postures this is marked in the Appendix. The embedded or dependent clause has the same posture as the main clause where this is not differentiated in Appendix B.

10.21 Applying posture to the text sample.

Discourse signals of posture may involve those elements of a sentence that are not part of repetitive lexical chains themselves but exert some influence upon how chains interact or are to be reinterpreted. For example, in text CCP, formulations such as *In comparison... Two conditions were considered...in vivo...after administration...* do not enter into long range cohesive chains (i.e. they may occur elsewhere but are not intuitively considered as identity chains: Halliday and Hasan 1989). Instead, they play an organising role in postures:

\$3 Two experimental conditions were considered: exponentially growing cells and plateau-phase cells...
\$8 In comparison, the Ci/Ce for etanidazole (ETA)... remained approximately constant at I for all values of melanin contents.
\$9 Treatment of Na1+ tumours in vivo with L3HIPIMO resulted in a tumour:blood ratio of about 3 at 30-60 min after administration.

In \$3, the reporting verb in *were considered* functions as advance signalling together with the prospective signal that there is going to be a list of two items (*Two experimental conditions*).

In \$8, the phrase *In comparison....* operates as a conjunctive discourse signal relating a previously stated proposition with the rest of the sentence, taking the discussion away from testing all conditions on one drug (*PIMO*) to another (*ETA*). \$8 is therefore an explicit logical encapsulation. Sinclair does not rule out multiple postures, and so \$8 can also be seen as an example of a retrospective verbal echo, where *treatment of* and *resulted in* echo experiments previously formulated in \$5 by *cells were exposed* and by four separate report statements (*ranged from* \$5,\$6, *this increase* \$7, *remained constant* \$8). We have given explicit encapsulation priority in our text sample. However, in cases where prospection and encapsulation occur together, prospection has priority. Sinclair gives an example of elliptical encapsulation being used in a sentence that also prospects a new topic: *The Prince of Wales* [Topic selection] *is among those who think it is high time they should.* [Ellipsis].

In \$9, there is no explicit clue to the sentence's relation to the previous argument. The items *resulted in* and *after administration* are new items, even though they may be seen to paraphrase previous experimentation (a case of a similarity chain, Hasan (1989)). Instead we claim that this is implicit logical encapsulation, where the encapsulation of experimentation is not explicitly signalled. Sinclair finds this type of relation prevalent in

his analysis of a journalistic essay. In cancer research articles, as can be seen in the analysis below, rephrasing encapsulation becomes more prevalent, although implicit logical encapsulation has a major role to play in certain rhetorical sections, especially Methods sections.

Before we summarise the main results of posture analysis in the ten sample texts, table 8 below sets out the number of posture types found in the whole sample and tables 9 to 14 detail posture for each rhetorical section:

Table 8: Postures in pharmaceutical research articles.

Encapsulation	79%	implicit	16%
		explicit	13%
		refocussing	26%
		rephrasing	21%
		including	3%
Prospection	6%	attribution	2%
		advance signalling	3%
		topic selection	3%
Verbal Echo	12%		
Overlay	4%		

Table 9: Posture in ABSTRACT sections

Posture type	Field			
	Structural Chemistry	*Cancer Research & Biochemistry*	*Total*	*Subgenre %*
Encapsulation				
-implicit	3	5	=8	15%
-explicit	1	8	=9	17%
-refocussing	**5**	**11**	**=16**	**30%**
-rephrasing	5	7	=12	23%
-including	0	2	=2	4%
Prospection				
- attribution	0	0	=0	
- advance signalling	0	0	=0	
- topic selection	0	0	=0	
Verbal echo	2	2	=4	8%
Overlay	1	1	=2	4%

Table 10: Posture in INTRODUCTION sections

Posture type	Field			
	Structural Chemistry	*Cancer Research & Biochemistry*	*Total*	*Subgenre %*
Encapsulation				
-implicit	14	3	=17	14%
-explicit	2	14	=16	13%
-refocussing	**19**	**16**	**=35**	**29%**
-rephrasing	13	10	=23	19%
-including	3	2	=5	4%
Prospection				
- attribution	3	3	=6	5%
- advance signalling	2	2	=4	3%
- topic selection	0	3	=3	2%
Verbal echo	6	5	=11	9%
Overlay	1	1	=2	2%

Table 11: Posture in METHODS sections

Posture type	Field			
	Structural Chemistry	*Cancer Research & Biochemistry*	*Total*	*Subgenre %*
Encapsulation				
-implicit	**12**	**31**	**=43**	**32%**
-explicit	2	1	=3	2%
-refocussing	5	32	=37	28%
-rephrasing	14	6	=20	15%
-including	0	0	=0	
Prospection				
- attribution	0	1	=1	1%
- advance signalling	0	0	=0	
- topic selection	4	3	=7	5%
Verbal echo	8	10	=18	13%
Overlay	2	3	=5	4%

Table 12: posture in RESULTS sections

Posture type	Field			
	Structural Chemistry	*Cancer Research & Biochemistry*	*Total*	*Subgenre %*
Encapsulation				
-implicit	3	11	=14	10%
-explicit	8	16	=24	18%
-refocussing	**7**	**20**	**=27**	**20%**
-rephrasing	6	18	=24	18%
-including	0	6	=6	4%
Prospection				
- attribution	0	5	=5	4%
- advance signalling	1	2	=3	2%
- topic selection	1	3	=4	3%
Verbal echo	**7**	**16**	**=23**	**17%**
Overlay	2	3	=5	4%

Table 13: Posture in the RESULTS-DISCUSSION sections

Posture type	Field			
	Structural Chemistry	*Cancer Research & Biochemistry*	*Total*	*Subgenre %*
Encapsulation				
-implicit	14	10	=24	13%
-explicit	13	13	=26	15%
-refocussing	15	20	=35	20%
-lexical	**16**	**22**	**=38**	**21%**
-including	5	3	=8	4%
Prospection				
- attribution	4	3	=7	4%
- advance signalling	0	0	=0	
- topic selection	4	2	=6	3%
Verbal echo	**15**	**15**	**=30**	**17%**
Overlay	3	2	=5	3%

Table 14: Posture in the DISCUSSION section

Posture type	Field			
	Structural Chemistry	*Cancer Research & Biochemistry*	*Total*	*Subgenre %*
Encapsulation				
-implicit	4	3	=7	12%
-explicit	4	4	=8	14%
-refocussing	4	11	=15	26%
-rephrasing	**8**	**9**	**=17**	**30%**
-including	0	0	=0	
Prospection				
- attribution	0	0	=0	
- advance signalling	0	0	=0	
- topic selection	2	0	=2	4%
Verbal echo	2	3	=5	9%
Overlay	0	3	=3	5%

10.22 Quantitative summary of postures in the text sample.

Posture is not a model that has been applied to many texts and it would be wrong to suppose that the proportions we find above indicate properties of this genre that are different to other types of discourse. Other text types do however appear to have a very different distribution , and my own anaylsis of a political speech suggests that verbal echo and overlay are dominant features of one example of such discourse (Gledhill forthcoming). Table 8 does however provide us with a baseline on which to judge variation within the sample, and that is the kind of data that we need in order to test the reformulation hypothesis. We can see from tables 9 to 14 above that the more frequent types of discourse signalling are unequally distributed across the text sample. Generally speaking, there is a move from selective refocussing to general rephrasing and a parallel cycle from explicit conjunctive signalling then to implicit relations and back to explicit lexical relations.

Let us assume that if a rhetorical section has a higher percentage of one type of posture than the text average, then it 'prefers' this type of signal over other, below-average posture types. To summarise the tables in words: Abstracts prefer explicit cohesive signals, especially refocussing and explicit conjunctive cohesion. Thus abstracts appear to combine

linear presentation of items and explicit rephrasing (typical of Discussions). Similarly, Introductions tend to use refocussing and they account for most of the instances of prospection in the sample. This is consistent with the Introduction's role of opening the research gap and predicting how the following text is to fill the gap. Methods sections on the other hand distinguish themselves from the rest of the sample by their singular lack of explicit signals, and this supports the intuition that Methods sections rely on experts' background knowledge for coherent reading. Conversely, Results sections prefer explicit signalling devices and (together with Results-discussion sections) the authors appear to use verbal echo as a means of reformulating data initially stated in Methods sections. This feature is discussed in the analysis below. Finally, Discussion sections account for the largest proportion of lexical rephrasing encapsulations in the text sample. This accords with our findings in Chapter 9, where rephrasing is associated with increasing use of grammatical metaphor towards the final stages of texts. This mechanism is also shared by Abstracts.

There is also evidence to suggest that some postures function in complementary pairs. Logical acts and deictic acts are constantly distributed throughout the sample (on average 24% : 40%). Since there is considerable movement within the categories themselves (in the Methods section in particular) this would suggest that implicit encapsulation coincides with refocussing encapsulation, and explicit encapsulation appears to coincide with rephrasing. Postures somehow act in conjunction, and this is borne out by the analysis below.

To summarise these patterns across the whole the research article (from Abstract to Discussion), refocussing can be seen to be a preferred discourse signal in early (Abstracts, Introductions and Methods) sections, verbal echo is preferred in later (Results and Results / Discussion) sections and rephrasing in Discussion sections. Verbal echo itself involves refocussing (where single items are repeated) as well as rephrasing where a set of refocussed items are reprocessed in relation to each other in colligation. But it clearly also has more text to echo by that point and this may be regardless of the rhetorical section. Nevertheless, we can conceive of the Results section as an intermediate section: it is the point at which explicit relations become more prevalent in the text sample and a point where, as we have noted, data are reformulated as results. In the detailed analysis below, we note that explicit cohesive signalling interacts with verbal echo and rephrasing (in many cases they co-occur, especially in Abstracts).

The global discourse signalling pattern is therefore:

A	I	M		R	RD		D
refocussing			=>	verbal echo		=>	rephrasing

A	I		M		R	RD	D
explicit		=>	implicit		=>	explicit	

These preliminary findings need to be supplemented. For example, it is necessary to indicate what changes in posture occur within rhetorical sections. Secondly, the weight of discourse signalling for a specific rhetorical section is unlikely to be homogenous for a set of texts, as we have seen in the analysis of grammatical metaphor. It is therefore necessary to concentrate on how discourse signals relate specifically to the construction of scientific claims rather than attempt to account for the rhetorical shifts in each text. In particular, we need to see whether discourse signals correlate with global writing genres such as 'report' or 'explanation'. The reformulation hypothesis claims that there is such a relation and this is tested in each text in the remaining sections of this chapter. We concentrate on one text: CC as a model of this kind of analysis, and report on the main findings for the other nine texts thereafter. The posture assignments for each sentence / clause complex for the text sample are set out in Appendix B (numbered from B1 to B10).

10.31 Posture in CC (Communication): a preliminary analysis.

As can be seen in Appendix B1 there is apparently no pattern to posture in text CC. CC is atypical in a number of ways; in particular, it is a 'communication' and the journal CC obliges authors to write one sentence abstracts. However its progression of postures are typical of the sample as a whole and we set them out below:

Table 15. Posture in a sample text from Chemical Communications.

$.	*Type*	*Relates to $.*
TITLE		
1	Text initial	
ABSTRACT		
2	E-Verbal echo	1
INTRODUCTION		
3ab	Text initial	
3b	E-implicit	3a
4a	E-refocussing	3b
4b	E-refocussing	3b
5a	E-refocussing	4a
5b/c	E-implicit	5a
6a	E-rephrasing	5a
6b	P-attribution	7a
7a	Prospected	
7b	E-implicit	7a
8	E-deictic-including	7b
9a/b	E-explicit	8
10a	E-implicit	9b
10b	E-refocussing	9b
11	E-verbal echo	10
12	E-implicit	
13	P-verbal echo	13a,14a
14a	Prospected	
14b	Prospected	
15	E-rephrasing	14
16a	E-implicit	15
16b	E-explicit	16a
METHODS		
17	E-verbal echo	15
18	E-logical	17
19	E-verbal echo	11
20	E-refocussing	19
21	E-explicit	18
22	E-verbal echo	20
23a/b	E-refocussing	22
24	E-rephrasing	23
25	E-rephrasing	24
26a	E-refocussing	23
26b	E-implicit	26a

$.	*Type.*	*Relates to $.*
RESULTS		
27a/b	P-advance labelling	28/29
28a/b	Prospected/E-explicit	
29a/b	Prospected/E-explicit	
30a/b	E-implicit	27
31a/b	E-explicit	30
32a/c	E-rephrasing	31
33a/b	E-refocussing	32
DISCUSSION		
34a/b	E/P-verbal echo	11 (etc.)
35	Prospected	
36a/c	E-refocussing	35
37a	E-rephrasing	36
37b	E-implicit	37a
38a/b	E-explicit/verbal echo	25 (etc.)
39	E-explicit	38
40	E-v.e. or overlay?	31?
41a/b	E-implicit	40
41c	E-deictic-including	41a/b
42	E-rephrasing	41a/b
42	P-topic selection	14b
43a	Prospected? /E-rephrasing	
43b	E-explicit	43a
44	P-topic selection	45
45a	Prospected	
45b	E-refocussing	45a
46a	E-explicit	45
46b	E-implicit	46a
47	E-implicit	46b
48	E-v.e. or overlay?	30
49	E-implicit	48
50a	E-verbal echo	47
50b	E-rephrasing	50a

Rather than global patterns suggested by the preliminary results above, we find series of micro-patterns that appear to cycle within each rhetorical section. As with other texts in the sample, it is these cycles which produce the global patterns we summarised above rather than neat linear patterns. CC's Introduction is typical: it progresses with a series of selective reformulating encapsulations, then a series of implicit and a series of refocussing encapsulations. Finally, as is typical in other introductions there are prospections about types of methodology to be used towards the end of the section. The Methods are also typical with little explicit signalling and a high degree of refocussing and a cycle of rephrasing-refocussing at the end. The Results have a mixed pattern as does the Discussion section (which is punctuated by above average prospection) although postures tend to occur in pairs, as they do elsewhere in the text.

CC is also typical of the other sample texts in that it stakes its scientific claim by stating that an observed chemical process can lead to new research avenues. We have already seen that grammatical metaphor enables this argumentation. By reassigning participant roles, the production of a chemical is gradually reworded as an abstract research idea (or cognitive process). At first CC rephrases experimental data as cognitive research processes in \$15 *This rationale led us to explore...* and \$16b...,*so that these ideas can be readily applied...* At a later stage, biochemical entities are also reformulated as empirical processes: '*the formation of 4-a... methylphosphonate*' is encapsulated by \$24 *this standard* and this is in turn encapsulated by \$25 *Other products formed were...* Later in the text we also find a biochemical reformulation of a result as a reaction: \$37 *This bioreversible protecting group...*, \$42 *In a related reaction...*, \$44 *An analogous reaction...* These items reformulate topical information (as standard terminology for compounds or a series of reactions) and represent a category of terms separate from anaphoric nouns (which label stretches of previous discourse). Lexical rephrasing therefore represents points in the text where metaphor is taking place and where claims are being reassigned. This does not show that claims reside in specific points in the text (i.e. in rephrasing cycles); it demonstrates rather a mechanism by which certain information is promoted to a different status within the text. Lexical reformulation represents an important mechanism of change in the status of the claim. We can also see that it represents the 'taxonomising' genre that Martin (Halliday and Martin 1993) has claimed as central to scientific discourse. We find that other discourse signals are involved in the complementary process of 'explanation'. With this in mind, we trace posture in the rest of the text sample with particular attention brought to the question of lexical reformulation.

10.32 Posture in JCPT9 (Experimental).

As can be seen in Appendix B2, the majority of postures in JCPT9 appear to be either implicit (E-logical) or deictic encapsulations which either repeat items from the previous sentence / clause or refer to it selectively (E-refocussing). Posture in the Abstract tends to move from refocussing encapsulation towards lexical rephrasing (*in all cases, these results*) and implicit logical encapsulation. In the Introduction, there is a similar progression, interrupted by prospection (attribution: enumeration: *one approach would be*) and then lexical rephrasing encapsulation (*these results, in the light of these data*) and one case of deictic reformulation: *this instability*. In the Results-Discussion section, which encapsulates cyclically as in CC above, lexical rephrasing reformulation becomes more prevalent at later stages. Around half of these instances are terminological rephrasings or superordinates where the rephrasing is also modified by some element of evaluation which relates to the status of the claim: (*the NMR spectra, this bioreversible protecting group, this moderate chemical stability, a similar reaction, this proposed mechanism, in a related reaction, the reactions, this degradation*). Other rephrasings are anaphoric references to results (*This result confirms that, this result suggests that, these results suggest that, in all cases this result*).

10.33 Posture in JCPT10 (Experimental).

JCPT10's Abstract is typical of the sample and contains only lexical rephrasing or selective encapsulation (*these observations, the results* and terminological reformulation: *these compounds, a related series*). The Introduction also finishes with lexical and selective reformulation (*such studies, in this paper*). We again see that rephrasing appears to be a standard 'end game' for most rhetorical sections. The Methodology section contains largely logical encapsulation, with a short stretch of reformulation (*further polarisation, these compounds*) and an anaphoric 'utterance': *the same argument*. The RD section has three cycles of deictic reformulation - at the beginning (reformulations of *compound, binding*), in the middle (*similar ring opening, this tertiary carbon atom, such reactivity*) and towards the end with rephrasing labels and reformulations: (*these results, the hydrophobic group, these compounds, this modelling procedure*). These research process reformulations can be seen to reformulate the empirical data as research findings. Again, this appears to be a characteristic of Discussion sections.

10.34 Posture in TL (Communication).

Text TL provides the simplest yet one of the most interesting cases of posture. TL's Introduction is a series of deictic selective and implicit logical encapsulation (again, typical of the sample). However, the RD section consists almost entirely of encapsulations based on deictic reformulation from \$12-17 (excepting \$13) (*the minor products, the ring-opening reaction, these alcohols, these acetates, the seleno compound*) and the concluding section \$20, 22,23 (*this route, this total synthesis, this methodology*). In the first stretch compounds are rephrased, while in the second each stage of the synthetic process is reformulated by a new superordinate item. As seen in Chapter 9, the claim of the text is that a new 'synthetic process' has been found, and as the biochemical methodology progresses the synthesis is resignalled gradually from a 'route' to a 'strategy'. This is in fact the only way the claim is explicitly signalled in the text: at the point (in \$23) where *JG* relates the new compound with drug therapy (a quality not mentioned since the introduction). Posture serves not just to signal but to progressively build the claim through the text. This discourse signalling is reflected also in grammatical metaphor and suggests that rephrasing reformulation and logogenesis are major organising features of the text.

10.35 Posture in JCPT3 (Experimental).

In JCPT3, a long string of refocussing connections links each sentence of the Introduction. We have seen that this is typical of Introductions in the text sample but given that rephrasing has been seen as a mechanism for reassigning claims, what role would refocussing have in the usual function of the introduction: to establish a research space? In the text we can see that refocussing does function to establish a claim, although now the reassignment of claim is done by the verbal process rather than the lexical reformulation seen in other sections. In JCPT3 each sentence brings a new perspective to the pharmaceutical application of PMEA by refocussing on the same biochemical entity: PMEA, but reformulating it in attributive or relational clauses. PMEA is related to evaluation in \$3 *further evaluation as a drug for the treatment of AIDS* and is equated with a biochemical inhibitor in \$4 *It is a potent and selective inhibitor of the human immunodeficiency virus*, and to activity in \$5 and in \$6*: PMEA has shown stronger in vivo antiretrovirus activity, PMEA is also active against a broad range of herpes viruses...* JCPT3's Introduction is seen to 'establish a research space' by empirical processes operating on a single unchanging entity. The claim has not been explicitly stated: the attributes of PMEA do however stand as an explanation, and we can see that the role of

refocussing and explicit discourse signalling in Introductions may be complementary to the 'taxonomising' genre which we have seen in rephrasing. When the time comes to reformulate this lexically, we assume that the additional relations and attributes of PMEA are 'textually' construed, that is understood. This happens in the RD section, which begins with implicit encapsulation, while the explanation of bond distances at the end correlates with a continuous stretch of deictic reformulation: *this structure, this torsion angle, the close 1,4 contact, the present structure.* This appears to be a typical pattern for our text sample: implicit encapsulation in the Methods section with increasing lexical reformulation in Results, Discussion sections and the final sentences of Abstracts and Introductions.

10.36 Posture in JMC (Experimental).

JMC's experimental section consists largely of implicit encapsulation and refocussing encapsulation, again typical of methods sections. JMC presents an interesting case of three Results-Discussion cycles. The *Crystal Structure Determination* section begins with implicit encapsulation and verbal echo of the prospected first sentence, and then ends with a chain of terminological rephrasing encapsulations (*this same thiazolidine conformation, this geometry, the crystallographic study, this type of hydrogen bonding*). The second RD section does not seme to have an overall pattern, except for the prevalence of refocussing encapsulation. The final RD section, *Comparison with sulbactam* begins with advance labelling (Tadros's *enumeration*) (*Several parameters were determined...*) and continues with either rephrasing (*these parameters, these figures, the principal differences, the hydrogen bonding described above, such a residue*) or explicit encapsulation (*consequently, it does not therefore appear that*).

10.37 Posture in BJ (IMRD text).

BJ's Abstract contains a typical stretch of refocussing encapsulation, followed by reformulation (*tumours of either type, this extra demand for glucose*) and ends with two including encapsulations (*this suggests that, this was supported by*) and an explicit encapsulation *thus*. The Introduction is dominated by encapsulating postures (mostly explicit or reformulations) with no example of implicit logical encapsulation. Four Methods subsections present typical implicit patterns, with a number of refocussing encapsulations, as well as verbal echoes ($35,38,45,47,55,58,59) where no direct link is made with preceding sentences and yet three or more lexical references coincide with previous sentences in the text (many echoing a sentence in a previous Methods section). The Results

section is predominantly made up of refocussing postures with several deictic lexical reformulations towards the end. There are a larger number of verbal echoes than other texts ($62,63,76, 81) in the first half of the Results section. The Discussion section contains five examples of verbal echo ($101,102,1113,119,120) and more explicit linking than in other sections. BJ's pattern of posture and reformulation is entirely typical of the overall conclusions we arrived at above.

10.38 Posture in JNCI (IMRD text).

As in previous texts, the Abstract in JNCI ends with deictic reformulation and lexical encapsulation (*similar lipolytic material, these findings*). The Introduction contains a long sequence of sentences linked explicitly, followed by a series of deictic specific and lexical encapsulations. The Methods section contains more implicit encapsulation, with towards the end verbal echoes referring to statements at setting out the basic methodology beforehand ($35->$33, $40-> $34, $40->$22). The Results section also contains several instances of verbal echo, referring to the Methods section, for example $41 *lipolytic activity from the MAC 16 tumor has been further characterized by DEAE-cellulose chromatography...* echoing $26 *The ...crude tumor extracts... were fractionated by anion exchange chromatography by use of a DEAEcellulose column....* Towards the end of the Results section, taxonomic report and reformulation become prevalent again (*These results suggest, structural elucidation of the cachexia-related lipolytic factor, The activity is, This result suggests*). In the Discussion section, verbal echo (on occasions echoing the Results section) and deictic reformulation (*such activity, such aged active ingredients, the activity of this material, another lipolytic factor*) begin the section, and explicit (*however, although, moreover, thus*) and lexical reformulation (*these data, this increase*) become more frequent towards the end. Again, this appears to be a typical pattern in the textual processing of scientific claims.

10.39 Posture in TPS (Review article).

TPS's Abstract commences with refocussing and explicit encapsulation and finishes with inclusive and reformulating encapsulation (*This suggests, such factors*). In the Introduction, deictic selective encapsulation predominates, with deictic reformulation involving nominalisation (*one possible mechanism for this stimulation*) As noted before, TPS is a 'review article', and its Methodology section is limited to one statement in $24, prospected by a question. The Results-Discussion section consists mostly of refocussing

encapsulation with two examples of prospection (both advance signalling *A second approach has been... Two observations support this hypothesis*). There are several instances in TPS where 'overlay' may be assigned. Abstract $6 provides a model for several sentences in the text (as in text BJ), and we can see that in the notions of *weight loss* and *nutrition / catabolism* are associated with the effects of *cachexia*. Long range verbal echo appears to be an important part of the textual cohesion of the text, although the posture model may occlude many instances of it where other postures (prospection and explicit signalling) take precedence.

10.310 Posture in CCP (IMRD text).

As with other abstracts in the sample, CCP's Abstract begins with a string of refocussing encapsulation and later has two examples of deictic reformulation (*this increase in the cellular uptake of PIMO, this high level of label*). The Methods and Materials section contains a typical number of implicit and refocussing encapsulations. In contrast with the Methods, the Results section contains only one example of implicit encapsulation, consisting instead of four verbal echoes referring at times back to the Methods section ($69 - 67, 70-43, 115-73, 124-120), explicit encapsulations (*however x2, in contrast x3, furthermore*) and deictic (anaphoric reporting) encapsulations (*in these experiments, these results, no similar increase*). The Discussion section is largely made up of refocussing and deictic reformulation encapsulation (*an association between PIMO and melanin, this localisation, these dying cells*) and deictic report (*the present in vitro results show that, on the basis of our results, the in vitro results also show that the latter findings*).

10.4 Posture and claims.

In addition to the overall results stated at the beginning of this chapter, the following findings emerge from detailed analysis of discourse signalling and posture in the text sample:

1) Lexical rephrasing tends to occur in the final moves of all rhetorical sections and overall in the final rhetorical sections of pharmaceutical research articles. Lexical rephrasing is associated with grammatical metaphor and with the reassignment of claims according to:
i) the status of the claim (expressed by anaphoric reference and conversion of concepts to research processes)
ii) position of importance in a technical hierarchy (expressed by terminological rephrasing).

In both cases rephrasing corresponds to Martin's (op. cit.) taxonomising report genre.

2) Refocussing encapsulation tends to occur in the initial moves of all rhetorical sections and overall at the initial rhetorical sections of pharmaceutical research articles. These kinds of posture are associated with the justification of the topic of a research claim:
i) by relational processes (equating an identified entity with another)
ii) by attributive processes (by attaching attributes to the entity)

In these cases also, refocussing can be seen as equivalent to 'packing' in grammatical metaphor and to correspond to parts of the text that constitute Martin's (op.cit.) 'explanatory' genre.

3) Other types of discourse signal have a less marked distribution, although we have noted above that implicit logical acts are used in conjunction with refocussing and explicit signalling does coincide with rephrasing. This would suggest that rather than being 'used in conjunction' these discourse signals have similar functions in terms of textual argumentation.

CHAPTER ELEVEN: Phraseology in the Pharmaceutical Sciences Corpus.

The professional context and internal composition of the Pharmaceutical Sciences Corpus (PSC) have been set out in the Data Collection section of this thesis. Chapters 9 and 10 set out the linear linguistic properties of a sample of the corpus. In this chapter, we observe the main phraseological and collocational properties of the corpus with a view to answering the phraseological hypothesis:

The phraseology hypothesis: *Collocational patterns correspond to rhetorical functions, and collocational patterns are consistent within rhetorical sections of cancer research articles.*

This chapter therefore describes the particular phraseologies of the corpus subgenres (the rhetorical sections of Title, Abstract, Introduction, Methods, Results and Discussion).

11.1 Salient items in rhetorical sections.

A keyword listing of all the words in a subcorpus provides us with a list of salient items that are of mixed frequency in the PSC corpus. These items can be sorted according to three criteria:

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

In the data collection section, we argued that grammatical items give the optimum amount of phraseological information for a medium-to-small sized specific corpus such as the PSC. Statistically the PSC is too small to provide interesting phraseological data for low frequency items (criterion 1) and the kind of data for criteria 1 and 2 would be more suitable for a lexicographic or terminological survey than a phraseological one. On the other hand, few phraseological studies have concentrated on grammatical items (criterion 3), because the amounts of data to be analysed are too large. Ironically, these studies are also too large to provide insights about specific genres. The idiom principle suggests that a phraseological unit will contain at least one grammatical item. If grammatical items are analysed first, then it follows that any lexical items of interest should emerge as organising elements of

phraseology. In other words an analysis of phraseology from the basis of grammatical items minimises the amount of data analysis needed by characterising global patterns first. Since grammatical items are more frequent, it is likely that any patterns they display will be more statistically interesting that those of lower frequency lexical items.

As detailed in Chapter 8.5 (Data collection), salient items are selected from each rhetorical section because they are statisitically 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. Salient items that enjoy a higher rank in Cobuild than in the PSC corpus are marked in **bold**. (Statistics for each section are provided later. only five grammatical items are salient in titles):

Table 16. Salient grammatical items in the PSC rhetorical sections.

TITLE	*ABS*	*INTRO*	*METHODS*	*RESULTS*	*DISC*
1 of	**but**	**been**	were	**no**	**that**
2 for	these	**has**	was	in	**be**
3 **on**	of	**have**	at	**did**	may
4 and	**there**	**is**	**then**	**not**	**is**
5 in	in	such	for	**had**	**our**
6 -	was	**can**	each	after	in
7 -	**that**	**it**	and	**there**	**not**
8 -	**did**	**we**	from	**the**	**this**
9 -	**who**	of	after	**when**	**we**
10 -	both	**to**	with	**all**	**have**

It can be seen that some sections are more 'Cobuild-like' than others. It is perhaps strange that 31 of the 55 words we analyse are in fact more frequent in Cobuild 1987 than in the PSC corpus. 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 PSC corpus. On the other hand salient items that are more frequent in the PSC corpus would have patterns which move the corpus as a whole away from the general language. In other words, when we analyse grammatical items as a whole, we characterise 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.

These subcorpus salient items (with the data that motivate their selection and phraseological summaries) are set out separately for each rhetorical section, below. We have attempted to limit the number of examples of collocation to five, although there is some variation in this. With long examples we have sometimes had to just include the head of complex nominals or miss adjuncts where they were not felt to be integral to the phraseology.

11.2 Transitivity processes and phraseology.

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

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

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

11.3 Phraseology in PSC 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 (58 332words) as detailed in Chapter 8. However, the items we analyse in the control corpus are determined by what is salient in PSC titles. A comparision can be worked out for the PSC corpus, but this reveals only five grammatical items with any salience. The Wordlist programme gives the following data (in the same format as discussed in Chapter 8 section 8.232):

Table 17: Title salient grammatical items from the Wordlist program

RANK	WORD	PSC Titles Freq. in subcorpus	%	PSC Freq. in whole 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 strikingly similar data for grammatical items: *of, on, and, in, by, via, its* and the marginally grammatical *self*. We analyse only the five most salient grammatical items in titles because the numbers involved in the PSC are too small for any statistical significance or to provide us with enough concordance lines.

11.31 Title salient item 1: Of.

In the *Medline* and PSC title corpora, *of* is the most significant salient item. 'Of' also eclipses 'the' in an Astec comparison with the Cobuild corpus, and is a salient item in the abstract and introduction sections, thus marking its phraseology as particularly typical of cancer research articles. 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 terminology. In titles we find no examples of quantification(a number of), or support (a group of). 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). We have divided the majority of left-collocates of 'of' into four groups of patterns.

Research processes are the most frequent in title expressions and typical expressions from *Medline* include nominal research process titles premodified by a topic-specific classifier with no article and post-modified by illness-related items most often involving cancer patients. The expression -study of- is typical:

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

The research process expression -evaluation of- (x15 in *Medline*) is different in that it is never premodified in titles (and is 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
factors aggravating postoperative recovery
factors affecting success of chemotherapy
factors affecting laboratory data
quality of life in postchemotherapy

The expression 'Evaluation of factors (X)ing Y' is more fixed than 'Case control study of (illness) Y', and is thus considered to be idiomatic according to the criteria we set out above. The semantic specificity associated with a small change of expression is a common feature of collocation throughout the corpus. 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 illness-related items introduced by -study of-:

A (research process) study on

clinical prediction
effects of continued...infusion
effectiveness of UFT against cancer
the inhibition effect of granisteron on...
usefulness of bleomycin in comparison with...

Our claim is that whole expressions starting with the most stable elements are involved in signalling a phraseological opposition 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. The distinction cannot be put down to lexical selection (or 'lexical projection' as in Universal Grammar (Cook 19), since both expressions share the same item. Similarly, phraseology is not dependent on the preposition. 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-:

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

-Management of- differs little in this pattern, except that the phrase is only premodified by one term, forming a terminological idiom: physiological management of (*invasive bladder cancer, terminal cancer pain, breast cancer patients in a group, brain metastases, localised prostrate cancer*).

Of the clinical 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		on
biphasic		aspirin	colorectal cancer
inhibition		surgical intervention	pancreatic cancer
		chemotherapy	metastases
prognostic		optimism	cancer related stress
therapeutic		somostratin	the growth of... cancer

This kind of pattern is a 'collocational framework' (Sinclair and Renouf's term 1991) and

can be seen to be similar in semantics to -study on- which 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 PSC corpus. What is interesting about them is that phrases such as -effects of- appear to be implicit in the longer chains, or are reformulated.

In the idiom -A case of- we find an example of a collocational cascade that is more oblique but clearly identifiable. While -case- is involved in the idiom '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
advanced oesophageal carcinoma treated by...
lung cancer responding significantly to...
pulmonary carcinoma which responded to treatment with
drug induced pneumonitis caused by oral etoposide.

11.32 Title salient item 2: For.

'For' is a significant salient item in the title and methods sections and generally signals a specific research problem, usually disease. It is used to postmodify complex nominals rather than in a prepositional phrase functioning as adjunct. In Titles it has the phraseological pattern: (treatment related item X) for (disease related item Y) which 'branches' between empirical and clinical process items:

empirical item:		disease:
consequences, estimates	for	colorectal / breast
implications, risk		advanced ovarian
risk factor		... cancer
clinical item:	for	cancer of the liver...
diagnosis, radiotherapy, resection		
chemotherapy, screening, therapy		
surgery, uretoscopy		

The concordance of 'for' also reveals a longer idiom with a different structure '(carcinogenic item Y) as a risk factor for (cancer Y)':

tobacco	as a risk factor for	lung cancer
ethanol		carcinogen free radicals
coffee		LOH mobilisation
tranquiliser		oncogene expression
perineal talc application		malignant melanoma

11.33 Title salient item 3: On.

'On' occurs in expressions that 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' (disease related items):

Research processes:		Empirical processes:
a retrospective study	on	effect
Basic study	on	influence
Clinical study	on	impact

In conjunction with these items 'on' is less involved in complex nominals than 'of' and 'for'. Its position in relation to these items should determine whether it introduces a prepositional phrase which functions as adjunct or as (nominal) qualifier. But in the following examples, the syntactic difference between qualifier and adjunct is blurred, especially when 'effect' is seen to be introducing both phrases:

#1 The effect of surgical intervention and neck cancer on whole salivary flow. (Qualifier of *effect* or of *surgical intervention and neck cancer*?)
#2 Blood transfusion does not have adverse effect *on survival after operations for colorectal cancer. A pilot study.* (Adjunct after on or after for?)

In #1, it is hard to say whether 'effect' or 'intervention and neck cancer' is head, and therefore which introduces 'on'. If 'effect' is seen to introduce 'on' then a collocational relation appears to be valid across functional and formal grammatical classes. The proximity of effect and on in #2 suggests that 'on' may introduce a qualifying phrase, in which case 'for' is candidate for introducing the adjunct. But the whole phrase after 'on' is mobile, and therefore syntactically speaking an adjunct. 'On' is also a key element in a fixed expression with a unique phraseology (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

11.34 Title salient item 4: And.

Conjunctive items are perhaps the least likely candidates to display collocational properties. Yet 'and' appears in a number of idiomatic expressions, the most marked being the following fixed expression: *combined* (research process 1 / clinical process 1) *and* (research process 2 / clinical process 2):

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

While 'and' is treated by the Cobuild dictionary as a conjunction that lists similar nominal groups, in *Medline* titles *and* is primarily used to list items that may be construed to be new combinations of cause and effect worthy of scientific enquiry: -(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: -(empirical process)+ (*between*) (disease related phenomenon) and (disease)-

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

although an alternative biochemical process version exists, as in: *[expression] differs between species and malignant tissues*

Besides relating previously unrelated items, the empirical item also introduces a listing of research / empirical process items that are corollaries (like the *salt and pepper...* idiom):

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

Similarly 'and' links complementary items belonging to a limted class of related items after the preposition 'in':

(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.

The collocational framework of complementary listed items also appears to be initiated by left-collocates of 'of' in expressions such as 'potential combination of X and Y'. This includes research and empirical process items: *detection, comparison, impact, role, effect, levels*.

11.35 Title salient item 5: In.

'In' is salient in four rhetorical sections in the corpus: this presents us with the opportunity to use 'in' to test whether phraseology is truly variable in the corpus, or just at variance with the general language. In fact, we find its use varies between certain sections. In addition, most of the PSC salient items are prepositions (in contrast to Cobuild salient items like 'that'), and this suggests that the research article genre differs from the general language at a basic grammatical level in areas such as prepositional and phrasal verb usage and construction and use of nominal groups. In titles 'in' functions in two ways:

1) as a prepositional phrase functioning as qualifier in complex nominals where the left collocate is a biochemical process. Where the head of the left collocate phrase is not the left collocate, the head item is usually an empirical or clinical item These are noted in bold:

changes in distribution of	cancer in	human, liver [etc]
intake and risk of		children, primary care
improved detection of breast		group practice, women
determination of screening for		rats, Singapore,
surgical therapy of prostate		the elderly, aged patients
gene	expression in	scrotal contents
		breast CYP1A1
receptor gene		cancer
		colorectal cancer

growth	factors in	gastric carcinoma
prognostic		
Expression of trypsin and other		HB carcinoma
p53-like.., p53 **expression** and other		(Y) cancer
diethyl analogue	cell lines in	culture
growth-regulatory		a p53 pathway
human bladder cancer		protein
larger auxiliary	metastases in	obese women
colorectal adrenal		patients with (cancer)
breast cancer		meginoma
evaluation of...hepatic		patients
prediction of auxiliary lymph node		tumour-bearing animals

2) in a postmodifying prepositional phrase where the left collocate is an empirical item whose statistical significance or medical potential is signalled:

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

Medical significance is also implicit in the following phrase:

bioreducible drugs and their	role in	cancer therapy

This second pattern is less prevalent in titles although there is an intermediate structure which includes a longer collocation involving the title salient item 'with'. With 16 instances of the *in patients with* collocation, we identify this as the most collocationally stable use of 'in' in titles. The structure is: (modified empirical item X) in patients with (disease Y):

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

In summary, the first pattern for 'in' suggests a general semantic tendency for the

qualifying phrase to specify the disease or the subjects in which the disease is to be found (the 'spatial' meaning), while the second pattern completes the semantics of the left collocate. The spatial use is not universal: in Abstracts where 'in' is also salient, the spatial use is not prevalent.

11.4 Phraseology in PSC Abstracts.

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

Table 18: Abstract salient grammatical items from the Wordlist program

RANK	WORD	PSCAbstracts Freq. in subcorpus	%	PSC Freq. in whole 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

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

11.41 Abstract salient item 1: But

The very high significance of but (compared with other grammatical items in abstracts) suggests that the reporting of negative results is a fundamental characteristic of abstracts. On can assume that positive results are announed 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 or to state negative empirical processes rather than quantify them (*however...X did not correspond, although this did not result in...*). To summarise, in abstracts negative data is quantified whereas in results sections negative data can be seen to be 'qualified'.

11.42 Abstract salient item 2:These

As we have seen in Chapter 10, 'this' functions to signal a refocussing and rephrasing reformulation. This function is shared by Discussion sections and a more detailed analysis is seen in our discussion of 'this' in section 11.7. We note here that 'these' differs from 'this' (in discussion sections) in that almost half of the occurrences of *these* are as pronouns introduced by *of*, while 'this' is mostly a determiner. The 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 below). This indicates a correlation with our earlier finding that abstracts tend to favour the use of deictic refocussing encapsulation. The high significance of *these* (according to Appendix C2) here also coincides with Nwogu and Bloor's (1991) observation that abstracts tend to employ simple thematic progression, linearly converting rheme to theme.

11.43 Abstract salient item 3:Of

In the control corpus of titles (as seen above), *of* was seen to play a key role in nominal groups with a typical treatment-*of*-disease pattern. Such a symmetrical solution-problem pattern is expanded in the abstract, the major difference being that while items in the title corpus tend to predict *of* with no strong right-collocates, in the abstract there are just as many significant right-collocates, such as *human, these, was*. Another difference 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, in Sinclair's terms (1991:82-83) as 'focus' rather than support.

The 'treatment-of-disease' pattern can be seen as an overriding pattern, but within this there is considerable phraseological change. We have identified four different problem-solution patterns of complex stereotypical phraseology with *of* for some of the most frequent left-collocates of *of* in the Abstract: (*effect, loss, number, presence*) and there does not seem to be any evidence to suggest that any such middle frequency item (often termed sub-technical items: Francis 1993) shares the same phraseology as any other. In particular, the solution-problem / treatment- disease pattern seen in the title does not appear to be fixed for each item in the abstract. For example, *presence of* has 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 an 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 most stereotypical patterns of the abstract. It is always preceded by some degree of measure or a methods-oriented specification of use (*daily, widespread, regular, intensive, combined, clinical, potential*) and followed by a specific drug X(1) and an expansion of the treatment and illness (*with drug X(2), in the study of illness Y, in the treatment of, in the evaluation of Y*) and finally followed by some degree of evaluation or a research process: *resulted in..., should be considered, is discouraged, is discussed.*

In a different kind of distribution, the significant collocate *loss* appears to have become

terminologised in the fixed expression *loss of heterozygocity*. *Loss* also appears in thematic position where a research statement is phrased in the passive or placed after the term (*loss of X...was found, occurred, occurring*), although there are reporting instances such as *suggest that* .... which 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 measurement-illness phrases mentioned above, the reporting verb precedes the expression (*shows/ confirms/ indicates ...the presence of, incidence of, absence of*). A fourth pattern is represented by the expression *number of* which is not immediately preceded or followed by a reporting discourse item. It may be that there is a differentiated pattern of phraseology in which *of* has a role as constructor of nominalisations of measurement and qualification (i.e. the first use mentioned above), in conjunction with expressions of research reporting and evaluation (the second use). The writer can thus choose to emphasise the 'self evidence' of the data by evoking phrases involving *number of*, or may wish to thematicise the study and be required to use stereotypical measurement-disease phrases, or alternatively thematicise the results and use an expression with items such as *effects*.

11.44 Abstract salient item 4:There.

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

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

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

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

11.45 Abstract salient item 5:In

'In' is used most frequently in three patterns:

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

In Abstracts, 'in' also introduces non-finite rankshifted clauses where given information on a chemical process is bundled in with the original information by explicative verbs such as *introduced, involved, implied* (as in: *this is a novel approach to adaptive resistance involved in the expression of ras oncogene*). In other sections, for example in titles, the most frequent use of 'in' is its spatial meaning (*in the liver, in cells*). In the Abstract this use is largely supplanted by a less specific meaning as in the use of *in* + *the* + (biochemical / clinical / empirical process), the most frequent of these involving the description of the mechanisms of carcinogenesis and tumour growth (*classification, suppression, treatment, transmission, dissemination, differentiation of the tumor, increase in the total number of cells*). On the other hand, *in* is followed by zero-article in the case of 'problem' items: cancers, subjects or specific disease-related entities (*cancer, breast cancer, tumor-bearing animals, patients, tumor-bearing mice, cytokines, methylene chloride*). It is likely that reference and other discoursal factors have a role to play in this distinction. But both these uses are of generic *the* in prepositional phrases and Master (1987) has claimed that discoursal factors (while crucial elsewhere) do not affect generic article / zero-article usage. So an alternative explanation may be that just as article usage is highly idiomatic in certain specific semantic domains in the general language, then it may be that phraseology becomes

more idiomatic in the specific language (as we attempt to demonstrate in this chapter).

11.46 Abstract salient item 6: Was

The simple past is the preferred tense for presenting the research article's present methodology and results. Ironically, as we have seen, the present is used to introduce previous research. This is contrary to previous research (Hanania and Akhtar 1985) and to Malcolm's (1987) distinction (past for generalisations, present for specific data). 'Was' reports the research article's (clinical) methodology and non-quantitative (empirical process) results. 'Was' in the abstract can be seen to play a completely different role to its present tense version: *is*. In the abstract, there are two patterns for *is*:

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

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

<u>was</u>	*metabolically expressed*
	immunologically reacted
	enzymatically deaminated
	induced
	carried

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

was (research process):

.... *observed, found, detected, determined, studied, seen, shown, investigated, demonstrated, performed, established, confirmed, compared.*

11.47 Abstract salient item 7:That

'That' as complement 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 introduce *indicate*,

values	indicate	that
findings	indicated	that
results		
information		

while *studies* and *results* also introduce *demonstrated.* A similar pattern is observed in discussion sections. One difference from the discussion section is the important rôle of 'that' functioning as relative pronoun in embedded clauses. It functions by referring 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*). Evidence of *that* (and, indeed *who*) as a salient item confirms Kretzenbacher's (1990) finding that embedded clauses an important characteristic of Abstracts, contributing to the traditionally 'compact' nature of the text at this point.

11.48 Abstract salient item 8: Did

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

The subjects of *did* reflect the typical sentence themes of the abstract: processes of tumour

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

11.49 Abstract salient item 9: Who

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

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

In particular, patients are never *given* drugs, they receive them (*who receive carboplasmin, receive Doxo, receive doxorubicin*). This is quite a clear example of the way phraseology helps to shape a specific view of 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 PSC corpus) allows us to establish its meaning within the corpus not only as more specific than 'personnel organisation' 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, fed, given...'. We did find, however, one instance of mice infelicitously 'taking part' in an experiment:

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

11.410 Abstract salient item 10: Both

Both signals a noun group complex, another possible characteritc of 'compaction' in Abstracts. In many of the cases where *both* is used as a linking conjunction, it is largely redundant. The following sentence is typical:

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

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

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

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

11.5 Phraseology in PSC Introductions sections.

The PSC introductions subcorpus contains 59 724 words. The Wordlist comparison with the PSC corpus gives the following data:

Table 19: Introduction salient grammatical items from the Wordlist program

RANK	WORD	PSCIntro 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

11.51 Introduction salient item 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). The passive perfect appears to polarise around a semantic difference between research process verb 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 (*in order of frequency >10: reported, shown, demonstrated, found, observed, identified, studied, described, obtained, published, conducted, detected, investigated*). We consider all of these research process verbs. However, this pattern also involves 3 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 as they are expressed they refer more to the research activity of the discourse community than to that of the authors. In this kind of phraseological distribution we refer to them as research utterances.

The verb 'shown' appears in both lists, and we demonstrate below that it has a different distribution to other verbs. Although each list has a consistent semantic content, we need to demonstrate that different lexical items are accompanied by specific variations on the global pattern. For each verb, we argue that these variations are consistent in themselves.

The first right-collocate of *been* with 40 occurrences is *reported* with the following phraseology: (biochemical process) have been reported to (projecting clause involving quantification):

p53 gene resistance	has been reported	to be very frequent
drug resistance		to be different in 2 case studies
antigen mechanisms		to be frequently carcinogenic
the LOH mechanism		to cause significant immunological damage
S-transferases		to produce metastasis in several species

A less frequent but similar phraseology involves *reported in* (+quantification):

gene inactivation	has been reported in	a number of cancers
MP substitution		a high percentage of carcinomas
LOH from 18q		several human cancers
low effects of inhibition		many tissues
drug resistance		mammals treated with PIMO

The second right-collocate in this pattern is *demonstrated*, with a similar phraseology but followed by adjuncts or embedded clauses relating to other biochemical processes rather than measurement:

serum	has / have been demonstrated	to be only aromatic
biodegradation		to form in vitro complexes
catabolites		in the original tumour
receptor myocyte		by trypsin formulation
endothelin recognition		by cloning tumours

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 we find a systematic difference where the expression relates the causes with tell-tale signs of cancer: (biochemical process) have been associated *with* (cancer Y):

Retroviruses	has/ have been associated with	hepatic cancer
Ras gene		specific neoplasia
high doses of toxin		gastrointestinal bleeding
mutation in these genes		haemic neoplasms
its effects on human health		the occurrence of cancer

A similar pattern is seen with *implicated* except that the pattern is: (biochemical process) have been *implicated in* (disease-related process Y) but the disease-related item is more specific than in the *associated with* pattern:

...implicated in...
regulating cell differentiation
in the development of cancer
the t-programming process

The third exceptional empirical item in the first pattern also 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			utilized....
This type of assay				
the macrolide technique				
A cross-characterisation technique				

Utilized is mostly interchangeable with *used* but is less frequent:

... have been used/ utilized to	study, evaluate, prepare...
... 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
	combination chemotherapy
	a recent synthesis
	the delivery of amines
	cancer therapy

The clauses introduced by the second (extraposed + research utterance) pattern have a less technical semantic scope than those in the first and generally express some empirical relational process. The embedded clauses can be seen to be past results framed in terms of a new (present tense) research direction but involving less technical or relational verbal processes (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.

A major claim of the phraseology hypothesis is that phraseological patterns are not due solely to phraseological 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'. 'Show' is used in both the passive perfect 'reporting' pattern and the extraposed 'research utterance' pattern but its use does not affect the overall phraseology: the first pattern involves biochemical technical processes, the second empirical semi-technical processes. In the first pattern (24 instances), the expression introduces biochemical processes as technically specific as those introduced by *demonstrated by* and *associated with*. This time, however, the postmodifying elements are non-finite clauses with verbs used semi- technically:

the disease	has been shown to	have a decreased resistance to
TNF alpha		efficiently deliver the toxicity of ricin A
a structural analogue of histidine		provoke an immune response
Quercetin, a lipoxygenase inhibitor		exhibit antitumour activity in vitro
encapsulation of dXR...		act as an in vitro inducer

There are only ten instances of *show* in the extraposed pattern, but they share the same phraseology as 'it has been established that'. The interesting difference is that 'it has been shown that' is preceded either by a temporal or adversative adjunct or subordinating conjunction and (if subordinated) is followed by a main clause justifying a different research approach or mitigating the old one with a negative:

It has previously been shown that a single dose of NM vax produces an inhibitory effect.
Recently it has been shown that Ras has potential protein-inhibiting properties...
However, it has been shown that use of this product can expose consumers...
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.
While it has been shown that these metabolites are frequently observed in breast cancer, their decline over time suggests that they are not a prerequisite.

These sentences are a clear case of phraseology extending 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 we assume that writers choose it to distance themselves from the possibly more subjective 'cognitive' verbs of the same phraseology. We propose that each

time a different verb is chosen, it has a different set of collocates (resulted in + measurement, for example), but that the phraseological paradigm assures that the new verb is interpreted in the light of the conventional expressions which share its phraseological pattern. The reason for the use of adjuncts may be that the semantics of the verb 'show' are not specific enough to suggest a temporal framework in which a wider discourse community is involved, which would be the case if the verb were 'proposed, concluded'.

11.52 Introduction-salient item 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 X of where X is a quality where 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 'has received' where 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.

The relational or possessive use of 'has' also involves almost obligatory quantification:

		Quantifier / Evaluation		Empirical / Biochemical item
granisteron	has		4	promoters
this compound			high	resolution power
the charged diester			lower	capacity
the inhibitor	has	a	profound	effect on its structure
the factor		a	peak	incidence between...
the disease		a	broad	spectrum of clinical indications

11.53 Introduction salient item 3: Have.

We have seen that perfective aspect together with extraposed expressions with 'has/have been' are the convention for reporting previous or 'given' research processes, while the present tense, as we see for the next item 'is', and the passive perfective are used to report 'given' biochemical facts. This is in accordance with previous research on tense (Heslot 1982, Salager-Meyer 1992). 55% of the instances of 'have' are involved in the expression 'have been', discussed above. Of the remaining instances we find one parallel expression with 'has received': 'have received (little, much) attention', but also 'have attracted (much, a lot of) debate, attention' and a similar distribution of empirical (evaluative) and measurement items after the relational use of the verb:

.... have	a	profound enabling effect
	a	good prognosis
	a	high glycolytic rate
	a	high prognosis potential
		poor capacity
		poor oral availability
		significant role
		totally different molecular framework
		well-documented effect

This is appears to be a regular phraseological unit: almost all items following a relational attributive / possessive 'has' or 'have' are premodified by explicit evaluation. Another striking regularity in usage is provided again by the verb 'show', this time used in the active 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 coworkers have shown that etc.). A similar and important use of the verb is introduced by 'we' (except that the general pattern is 'we have found that') but this is discussed below under 'we'.

11.54 Introduction salient item 4: Is.

Unlike the possessive relational process use of 'has /have', the role of 'is' is uniquely used for explicit evaluation rather than signalling identity which is its usual pattern in the general language, at least according to the Cobuild dictionary. The phraseological patterns are (in order of frequency):

1) It is (empirical item) that (treatment related item X) (biochemical / empirical process):

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

2) It is (empirical item) to (research process)

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

3) (Biochemical process) is (research utterance) to (biochemical process). There are only three possibilities for this type of expression, expressing 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

As a copular verb, as with the relational use of 'has, have', relational and attributive uses of 'is' involve explicit evaluation. This can be seen in equative relational clauses (occurrences >5 are underlined):

Specific biochemical process or item
Pancreaitis, resistance to therapy, BORA, the Winsford deposit... disease Y...

is a

Evaluative	*Empirical item*	*Evaluative*	*Biochemical item*
common	predictor	important	target
appealing	alternative method	effective	inhibitor
critical	parameter	potent	derivative
major	sign	potential	agent
imperfect	route	common	analog
complex	issue	strong	inhibitor
		rare	disease
		new	solid

Similarly, in existential clauses there is also almost always some explicitly evaluative element:

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

When 'is' is used in equative relational clauses, the element of evaluation is transferred to a notion of 'measure', as in the fixed expressions 'is one of the most...is one of the main causes of''. As seen in other areas of the corpus (especially surrounding the item 'of') disease- and treatment- related items have stereotypical patterns in attributive 'is' clauses. Only disease related items, for example can be 'associated with':

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

Conversely, only treatment related items can be 'more' (+ empirical property):

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

The reason for these patterns stems fairly straightforwardly from the research activity. Diseases are being associated with potential causes, while treatments are being compared and measured. So a phraseological pattern correlates according to some convention with the common semantic categories naturally involved in the research. In addition, collocational analysis of 'is' also reveals a limited set of items which can introduce nominal complement (projecting) clauses, and these items are almost always empirical and mostly 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, we find that the evaluative pattern is not prevalent in all uses of 'is'. In the introduction corpus, when the researchers are saying that something is *not* something else, explicit evaluation disappears:

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 above relates to empirical or research processes in similar expressions to the pattern 'Although it has not been shown that' described under 'been' above. To summarise, affirmative expressions with 'is' do function much more in terms of modality than simple equative expressions. 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.

11.55 Introduction salient item 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 alkylating carcinogenic other	agents	such as	NMU BCNV nitromidazoles TCPOB-08
use of hormonal several DNA metabolic detoxifying	enzymes	such as	dismutase exonuclease transferase acetates
	tumors such as		Wilm's melanoma maleic myeloma the adenocarcinoma 755 MCF-7

The reformulation appears to be bi-directional: the first item could be a new item, while the qualifying phrase '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 in use in technical writing. What is clear, however, is the function of rephrasing reformulation which confirms our initial findings (on posture) that this is a fundamental mechanism in report writing and explanation.

11.56 Introduction salient item 6: can.

'Can' expresses potential empirical procedures or biochemical processes. Two patterns emerge for the modal 'can', either in research oriented passive constructions or in active technical expressions:

1) (General clinical or empirical process) can be (research / empirical process-ed):

alterations	can be	prepared	applied
variants		deciphered	prevented
ideas		correlated	determined
methods		considered	classified
therapies		attributed	derived
products		obtained	

Some technical biochemical processes are also used in this expression: *transmitted, modulated, coupled, induced.*

2) (Specific biochemical process / item) can (technical biochemical process):

gene products	can	dimerize
cytokines		flip
IL-2		hydrolyse
differentiated cells		induce
gingivalis		undergo malignant transformation
DNA		metabolise
PMEA		inhibit

11.57 Introduction salient item 7: It.

Most of the uses of 'it' have been described in the discussion of 'it is' and 'it has been' + (research process) above. Other extraposed clauses include:

it was also	found that	the polymer was not stable
it was		it causes higher overall cell counts
it was		although stability outside the cell...

Other verbs that occur in this pattern are: *thought (x3), reasoned, reported, shown.* There are also a number of modal expressions such as: *it appears that*, and *it would seem that.* In parallel to other expressions of evaluative research utterances (it is essential to etc.), we find: *it would be worthwhile to.* 'It' is the most Cobuild-salient item 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 the absence of extraposed clauses is a characteristic of Introductions rather than the rest of PSC corpus.

11.58 Introduction salient item 8: We.

A rich variety of expressions are used when the research article writers present their own previous or current research. In many of the expressions involving the researchers as writers there are time expressions or deictic references to the writing process. These appear to vary according to the choice research process verb and circumstantial adjuncts:

Here	we compare	production in sheep expression of gene alpha spectra
In this study in the present in a subsequent	we examine	a combination of methods the activity of PKC (x2) the incidence of protein the distribution of PIMO the p53 protein

The combination of formulations is complex, but essentially we find the following recurrent elements (with variable positions): (time reference) (reference to this study / paper /report) we have (research process):

Previously (in this - a previous/ report - paper)
Recently (in this / a previous report / paper)
we have (in this study / previously, recently):

found that
investigated whether
investigated reactive effects
investigated other protonated exons
recently shown that
recently determined
previously reported (x3)
previously been studying
reported that mutant p53 causes
shown that
studied p53 expression
studied NAK cell susceptibility
studied tumour-drug distribution
succeeded in establishing ribosome-resistant cell lines
succeeded in catenating cyclodexrins
succeeded in establishing 2 ph1-positive ALL cell lines

A similar pattern uses the simple present tense, but this time used exclusively the verb 'report': we (time reference) (research process) (reference 'here') that (results):

we now report	that p53 overexpression is elevated in the presence of
	that epoxyalcohol also inhibits
	the results of our immunological studies
we report here	the results of a physical study
	the results of our study
	that 2DDP-subclones
we report	that growth in soft agar appears to involve.. substitution
	the synthesis of 3 substituted pyramidizole
	first isolation and characterisation
	characterisation of a new breast cancer cell line
	2 different approaches to synthesis

11.59 Introduction salient item 9: of.

"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. This further implies that collocation operates at longer boundaries than the phrase, an assumption that we saw was typical of the terminological view of collocation. 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: the aim / purpose of (this study) was to (+ research process) (measurable biochemical activity) (16 occurrences) :

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

The 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* were generally followed by some degree of evaluation or an empirical process (the effects of treatment X are demonstrated) here the phrase occurs as complement to 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 is 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 (hedged or negative research process):

The mechanism of action of human tumour model systems is
The mechanism of action of their cytostatic action appears to be 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 the Abstract's expression *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 premodifying term 'drug of choice' (6 occurrences) is also a frequent premodifier of *'in the treatment of'* and this would indicate the formation of a relatively stable expression. Even more striking is the level of reformulation of similar concepts for new drugs used in the following 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

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

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

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

11.510 Introduction salient item 10: To.

We have already seen the role of 'to' as complementiser in expressions with 'it is important to' and 'have been shown to be'; however this does not exhaust its role as complementiser in noun group projections in other salient expressions in Introductions. One particularly regular projecting clause takes the following form: (biochemical process: possessive) ability to (biochemical process):

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

In some cases, the phraseology reflects subject - verb patterns. 'able to', for example can either 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 major 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

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

response to DNA damage	leads to an arrest of the cells
This in turn	leads to increased conversion of the lactase
This process	leads to inhibition of intracellular concentrations
altered membrane transport	leads to degradation extracelluar matrix (ECM)the agonist 2-
methyl 5HT	leads to release of substance P

One rationale for this intriguing difference is that tense and aspect play a role in phraseology (we see elsewhere that it does for is/was was/have been) and that tense has a 'research orientation' meaning that has a more relevant role to play than 'real time', an observation which accords with Wingard's findings (1981).

We have already mentioned above (in discussion of 'that') 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*) in Abstracts and Discussions. This may reflect an increased use of indirect grammatical metaphor later on in the text. In Introductions we find mental research processes projecting explanatory clauses:

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

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: 'although (negative)'. We also note also that the negative which accompanies adversatives like 'Although' seems to operate in parallel with 'appears that' and comes either in the main or subordinate clause: it is as if the phraseology requires a negative expression but has no preference about where it is finally expressed. Again, one explanation for this variation may be that phraseology determines what grammatical choices are available with the final 'mechanism' of thematic choice and word order left to textual considerations.

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 phrasal prepositions: according to + research model:

according to in vitro criteria
according to soliton theory
according to the theory of Knudson (1985)
according to the mechanism we put forth.
according to tumor histology (Palmer et al. 1988)

and phrasal verbs, as with the very frequent compared to + biochemical process, or nominalisations of biochemical processes which take -to-, such as the equally frequent 'resistance to chemotherapy'. A longer phraseological unit emerges with the nominal 'exposure to': (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 that normally require 'to' very often involve 'cells' in a

complex nominal where the cells are related to another biochemical, often a reagent which in the case of cancer appears a lot of the time to be 'growth factors':

(Empirical /biochemical)		(Biochemical entity)
responses	of cells to	a wide variety of mitogenic growth factors
resistance		growth factors
susceptibility		hormones in growth factor
responsiveness		oestrogens
similarity		the antibody

The final point about phraseology in the introductions corpus comes back to the use of 'to' as complementiser. One of the more interesting patterns to emerge involves 'was' (the 4th highest collocate of 'to') where almost all of the expressions formulate the aims of the research paper. We have already seen 'This aim of this study was to'; however, the variety of expression we find using '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
The purpose of the current report was to generate and trap...
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...

Here the only permanent elements of the phraseology are the grammatical items 'was to', although the semantic pattern sticks very consistently: (research goal) was to (research process). The only exception to this seems to be where the aim is to 'do something', in other words 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. 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. In the case of Introductions, goals are presented as global research rather than the specific empirical or clinical processes.

11.6 Phraseology in PSC Methods sections.

The PSC Methods subcorpus contains 137 161 words. It includes Experimental and joint Methods - Results sections. The Wordlist comparison with the PSC corpus gives the following data:

Table 20: Methods salient grammatical items from the Wordlist program

RANK	WORD	PSCMethods 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

11.61 Methods salient item: Were.

As with 'been' in the introduction, 'were' is a significant marker of the passive. But whereas passives elsewhere in the corpus are research oriented ('have been identified' etc.) here the past passive (which is unique to the Methods section) is clinically or empirically oriented, involving sometimes highly technical verbs. In previous research Hanania and Akhtar (1985) found that the passive in Methods was found to be frequently a present passive which we do not find here, while Heslot (1982) and Wingard (1981) found that the simple past was prevalent in Methods sections, for which we do not find evidence in this corpus. In the literature, passive expressions have been classified in terms of the relationship between subject and verb (Sager et al. 1980, Heslot 1982, Hanania and Akhtar 1985, Swales 1990). We supplement this view of the data later, but for the moment we list some of the more frequent SV relationships, and we find that there appears to be a relationship between subject and the verbal process of the predicator:

anerobes	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 prepositions. We shall see later that these can be further sorted by verbal process. We term a 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) instruments / media:

Clinical process

were	analysed by	log rank test ANOVA test using analysis of variance	(statistical test)
	determined by	TLC scanner liquid scintillon counting the method of Chadwick et al. means of a Student's t-test the HPLC method	(clinical or instrument)
	killed by	cervical dislocation exsanguination CO2 anaesthesia CO2 asphyxiation	(clinical procedure)

	obtained by	measuring the fluorescence using a 1.5 mm diameter cork borer retro-orbital bleeding of mice injecting 3x105 cells into both flanks	(clinical procedure)
	prepared by	the reverse evaporation method the film method of Skoza et al. protein precipitation with acetonite dilution of the liposome dispersions	(biochemical method)

With 'for' (a Methods salient item) the passive construction is empirically oriented rather than clinical, with the following cascade:

Research / empirical process

were	analysed for	hormone traces significance	(observable item)
	calculated for	antibody depletion luteinizing hormone count	
	eligible for	the present study this study	(study)
	examined for	visceral defects malfunctions external defects	(disease-related item)
	used for	observation evaluation of patients the experiments	(research process)

With 'at' (another Methods salient item) 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 77 minute intervals 1 minute intervals
	incubated at	37 degrees C
	stood at	room temperature
	performed at	37 degrees C
	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 processes which have a variety of expressions. For example several idioms are used to express the (clinical or 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 or 'control groups'):

(animals) were	killed	by cervical dislocation
	sacrificed	by severing the dorsal aorta
	euthanized	after 82 weeks
	necrotized	by CO2 asphyxiation

Incidentally, we also find one instance of mice being 'shot', but fortunately (or perhaps, unfortunately) we assume that this means 'injected'.

11.62 Methods salient item 2: Was

While 'was' shares a similar passive phraseology with 'were', the difference between the two relies on the fact that items in the methods subcorpus tend to be groups of biochemical entities (cells, tissues, mice) while singular items in the methods subcorpus tend to be empirical process items used with biochemical processes such as :

GST activity	was expressed as
MPO activity	
Kinase activity	
reductase activity	

or research process items used with clinical process verbs:

DNA analysis	was performed with A FAS solution
Thus analysis	was performed using a one way analysis
Statistical analysis	was performed by the CELL fit method
retrospective analysis	was performed in a similar format

A particularly frequent pattern accompanies '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

Such regularity of expression suggests that certain phrases may be author-specific, it certainly suggests that some phraseological patterns are typical of a small number of texts in

the corpus, although it is difficult to say whether there is any indication of this elsewhere in the corpus.

Another difference in phraseology is that while 'were' expresses methodology by the expression '(biochemical entities) were (clinical process verb) by', singular items tend to have the following formulation: '(usually deictic) (empirical / research process) was (clinical / empirical process verb) using':

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 solvent was removed under reduced pressure (x5 instances).
the solution was run on the plates for the analysis (x5 instances).
the supernatent 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)

11.63 Methods salient item 3: At.

If the base meaning for 'by' is to signal research methodology in the methods subcorpus, 'at' in a similar way signals a category we have termed 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', often with the possible collocational framework of 'for (x hours) at (temperature x):

carried out	at 67 degrees C.
centrifuged	at 12 000 rpm
eluted	at a flow rate of
stirred for one hour	at room temperature
heated	at room temperature
incubated	at room temperature
maintained	at 72 + 30 F
measured	at 400mm
measured for 3 min.	at 37 degrees C.

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): (stirred) at (temp.) for (time.) until (empirical / clinical process item):

was / were	stirred at 20 degrees C. for 40 min.	until DNA extraction
		until processed
		until assayed
		until analysed

Some clinical verbs seem to occur with measurement items that are less specific, such as 'obtained at':

obtained	at a constant
	at successive treatment times
	at these ranges of azine ion
	at later time points
	at individual time points

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 expression 'at least' however also fits into the 'measurement' pattern:

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 prepositional use of 'at' for a certain place is not prevalent in the methods corpus, although we find instances such as: ***unidentifiable numbers are placed at the bottom of the***

scale.

11.64 Methods salient item 4: Then.

Cobuild-salient items like 'then' appear to function perfectly normally in the corpus when their uses are listed. We have seen however that the number of potential LGP uses the Cobuild dictionary ascribes to certain words (19 non-idiomatic uses for 'of', for instance) are usually highly restricted by the corpus. Despite being a very significantly 'Cobuild-salient' item, 'then' functions in a uniquely specific way in the corpus (in fact, it corresponds to one out of ten possibilities in Cobuild (1995 2nd ed.), and its position is more fixed). 'Then' appears exclusively as an adjunct before passivised verbs to signal a subsequent incremental step in the methodology. The most fixed phraseology involves the idiom 'added dropwise': 'the solution was added dropwise and the suspension was then heated' (x4 instances). The following clinical verbs are the 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

11.65 Methods salient item 5: For

We have seen above that a major use of 'for' is in a number of expressions to signal a very specific research goal for a stage of the analysis within the methodology. The kinds of research goal depend on the passivised clause, as seen above. In addition, a particularly regular phraseology emerges for the expression 'examined for' where the phraseology is: (animate donors / cells) <u>were examined for</u> (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

The phraseology is so regular that we assume that 'RT activity' in the following expression must be considered evidence of disease: *The heads were senally sectioned and examined for*

RT activity. A more direct expression of research aims consists of the expression 'used for':

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

The post-adjectival expression 'eligible for' is also classed as an empirical process because it bears on the relevance of certain data to the research :

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

11.66 Methods salient item 6: Each

Highly regular expressions involving empirical and clinical processes are clearly beneficial to the researchers in their 'indexical /reference' reading: there is simply no need for argumentation at this point in the text. The determiner 'each' reveals such a set of fixed expressions, and they are typical of many more such expressions in Methods and Experimental sections. The salience of 'each' is evidence that the language of this rhetorical section has been 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. We have already noted that implicit non-referential progression is typical of Methods sections, and the other prevalent pattern, deictic refocussing, also happens to be signalled by 'each'. 'Each' is used in a number of fixed 'measurement' expressions when the researchers want to emphasise the distribution and repetition of a series of clinical acts. Among the very fixed expressions we find three patterns:

Empirical quantification : dose

verified	at each dose level
entered	
repeated	
counted	
treated	

Empirical quantification: time interval

for	each	day
		month
		hydrolysis
		study
		rat

Empirical item: subject group

separated	from	each	colony
aspirated			mutant
removed			contact
prepared			treated region
withdrawn			sample

One example of the many fixed expressions we find is 'added to each well':

buffered saline was	added to each	well and incubated
11g of bromide was	added to each	antigen well
sample buffer was	added to each	well to dissolve.
5g of purified rabbit	added to each	well to dissolve the MTT formazan
3H leucine was	added to each	well to dissolve the sample.

11.67 Methods salient item 7: And

As with 'then' and 'each', the salience of 'and' is due to the sequence of methodologies being presented in the subcorpus. General clinical processes such as 'collected' are listed first, followed by more specific, technical clinical processes:

collected	and	concentrated
		exposed to methylene chloride
		incubated
		mixed
		radioactively determined
		treated
		stained
		(re-)suspended

or followed by general processes of location pending further experimentation:

collected	and	counterstored
		mounted
		placed
		stored

As with 'both' in the abstract, 'and' generally reveals nominal group complexes that we

might consider to be complementary:
adenomas and carcinomas (x8 instances)
amplification and sequencing conditions (x4)
forestomach and lungs (x4)

Some clinical processes appear to be equally 'inseparable', especially involving the methodological technique of 'staining' or processes that either always come first in the sequence or follow:

sected and stained with...
treated and counterstained with
removed and routinely stained with...
developed and stained...

cut and stained (x5)
cut and mounted
cut and plated
cultured and plated (x3)

11.68 Methods salient item 8: From

'From' reveals a preoccupation in the Methods sections with the source of data samples, particularly from organisms. We did not detail the use of 'from' after passivised verbs in the discussion of 'were / was', but the majority of examples here conform to a similar phraseology. 'From' also reveals embedded passive clauses in complex nominals (the 'reduced-relative' pattern). In the case of 'from' the basic semantics involve either a clinical verb (extraction of a biochemical from material sources), or an empirical verb (basing data on a specific methodology). The most frequent use is the meaning 'extracted': *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	extracted from PALL cells	
mRNA	was extracted from	the parent cells
tRNA	was extracted from	the exponentially growing cells

On the other hand, for the reduced relative expression 'obtained from' we find greater variety of expression. We also find both the clinical 'extraction from biochemical entity' as well as the empirical 'based on this data source' phraseologies:

Research data source:

cells	obtained from	Dr JH van Dierendonk
data		the above reaction
cultures		Sigma Chemical Co.
tissues		hospital recalls
values		the previous study

Clinical extraction:

DNA	obtained from	patients
cell lines		platelet rich plasma
mice		breeding colonies
tumours		control mice
A factor		green tea leaves

A clear example of an empirical phrase which only involves the 'data source' meaning is 'calculated from':

functions	calculated from	the bootstrap samples
intervals		data
X' serum concentrations		dose-response curves
size of p52 mRNA species		equations
second-order rate constants		a standard curve

'From' in qualifying 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

We also find the complex nominal phrase: (specific body location) tissue from (donor):

normal breast	tissue from	10 patients
spleen	tissue from	normal chinook salmon
embedded kidney	tissue from	10 control and 10 exposed animals
fixed	tissue from	the orbit of the eye
recipient	tissue from	24 cancer and 8 pancreitis patients

11.69 Methods salient item 9: After.

The phraseology of 'after' has been mentioned in conjunction with passivised verbs such as 'obtained, added, killed' (its 3 most frequent lexical left collocates). As a postnominal qualifier, 'after' is preceded by a time expression where 'after' introduces a nominalisation of a clinical process. In other words, the methodological procedure is presented in reverse order. This is different to its use in the general language, where it often introduces a time expression (*after two days, after a while*: according to Cobuild). Some typical examples include:

Clinical process.		Clinical nominalisation
were added 24 hours	after	amputation
determined 26 days		implantation
were killed 26-30 days		injection
cell growth was analysed 5h		tumor transplantation
exposure at intervals up to 5h		treatment
cultures grown 3 hours		the start of chemotherapy
regimes administered several hours		heating at reflux
l-action was applied for 2 hours		drug administration
determined 100 min.		injection
repeated every ten minutes		grafting of the tumor

In many cases the time reference is the adjunct 'immediately':

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

A similar example involves the verb 'obtained' but there is generally no time reference:

Tissues were obtained	after	the addition was complete
		the addition of 0.5ml water
		the first dose of interplasmin
		the first injection
		the initial dose

As with other grammatical items, a small number of highly consistent collocational frameworks emerge, in particular the very long phraseological unit: [(disease-related item) / (empirical / clinical process) within (precise time reference) after the (ordinal) dose / injection]:

mice were killed	within 2 hours	after the last dose
loss ... of weight	within five hours	after the last dose
tumors began to appear	within 24 hours	after the first dose of injections given
deaths occurring	within five hours	after the first injection

Long stretches like this are not 'cascades' since they do not involve 'steps' of variable but semantically consistent members. Stepped cascades would involve a different phraseology in either the first of the second phrase such as at each + (level) contrasted with for each + (time unit). Cascades indicate divergent phrases, with a new semantic category introduced by each collocational part of the whole. Instead, complex collocational frameworks such as [within X hours after first / last (clinical process)] are convergent: their semantic parts vary within a limited set of alternatives. We may argue that these are cases of 'phraseological units' where a collocational framework interacts with fixed lexical items in a linear string.

11.610 Methods salient item 10: With.

We have already mentioned the significant role of 'with' as the most frequent right collocate of 'were' in methods sections. Whereas in titles 'with' is a salient item used to conjoin similar research processes, in the methods subcorpus it signals an instrument or 'medium' by which the clinical methodology is achieved. Even more specific phraseology than that discussed under 'were' can be found with certain verbs which all have a delimited set of possible instruments:

activated with (biochemical solution)

were activated with	ethanol
	an equal amount of saline
	a cell suspension
	the culture medium
	blank human plasma

incubated with (subject-derived serum)

were incubated with	a mouse monoclonal antibody
	monoclonal antibodies
	antimouse antiserum
	test sera
	antirat IgG mixture

stained with (colouring agent)

were stained with	10% ammonium sulphide
	Alcian blue stain
	brilliant crystal blue
	nitro-blue tetrazolium
	monoclonal antibody

treated with (quantity of measureable substance)
were treated with 2 parts of ammonium persulphate
indicated compounds
concentrations of 5% ammonium...
various concentrations of 8 chloro cAMP
various doses of TPA

11.7 Phraseology in PSC Results sections.

Table 21: Results salient grammatical items from the Wordlist program

RANK	WORD	PSCResults 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

11.71 Results salient item 1: No.

'No' is the most significant salient item in the Results section, and its role in signalling changes in the data is similar to the pattern 'but...' followed by negative in the abstract. 'No' functions uniquely as a negative determiner, a usage that is not listed among the 12 uses of the word in the Cobuild 1995 dictionary. Its most frequent use is in the long phrase 'there was no significant (difference / correlation) between the value/s.' (all collocates >10). We can class similar phraseologies by the final preposition in the phrase:

Empirical statement	Data shape		Biochemical / clinical
There was no significant	change	in	radiosensitivity
	difference		plating efficiency
	increase		hydrolysis
	change		the time course of efflux
	variation		food...consumption
	Empirical relation:		
	association	of	EST alpha with GST nu
	effect	of	vermapil on accumulation
	effect	on	reduction of tumor size

In Results sections, affirmative statements of this kind tend to be expressed in the present tense: this is discussed below under the item 'there'. We also find several instances of the passive form of this kind of phrase, with a similar split between empirical relation / data shape:

No significant	relationship	was found
	association	was observed
	association	was found between tumor grade and LH
	change	was seen
	difference	was observed during the time period
	correlation	was observed with respect to rewrite mRNA

Note that the passive is used with research process verbs rather than the clinical verbs observed earlier in the abstract and methods sections. A version of the passive form (limited to 'observed') also exists which emphasises the biochemical process:

	Biochemical / Empirical process	Research process
No significant	temperature dependence	was observed
	survival prolongation	was observed
	lesions	were observed
	tendency toward sustained release	was observed
	time effect within one group	was observed

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
	consistent	pattern across concentration
	detectable	difference in the incidence of
	strong	evidence for tumor development

The same may be said of all relational process verbs (including possessive 'have'):

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

The role of the negative to indicate changing data can be seen in the following passive examples where the expression is introduced by a conjunctive expression:

Conjunctive	Empirical / Research process
In contrast, no	clear trend was associated
In contrast, no	clear correlation could be found
In contrast, no	similar increase in radiosensitisation was observed
By contrast, no	necrosis factors were found to be present
However, no	allele loss was observed

The expression with 'obvious' is the most frequent:

This caused	no obvious	antitumor effect
experienced	no obvious	pattern between the two periods
while	no obvious	difference was observed
with	no obvious	effect

Other post-verbal uses of 'no' reveal the delexical nature of verbs used to report results in contrast to the empirical / research based verbs shown above:

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
screening	revealed	no	activity
postmortem 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. In all of the expressions, no indication is given to determine whether data increased or decreased, and this signals that in Results sections when researchers write about what didn't happen they talk not in terms of the shape of the data but in terms of explanation of empirical terms - that is, how close it comes to their hypotheses about cause and effect.

11.72 Results salient item 2: In.

'In' is used in three types of phrase in the results corpus. The first is to indicate positive results which usually involve a higher score or increased amount in terms of measurement. This can be contrasted with the negative results presented above, which we characterised as 'without direction', usually indicating only the relevance of the result to the empirical model. The second is closer to the essential spatial meaning of 'in', indicating where a specific biochemical process was found / observed in the bodies of patients or subjects. The third takes the form of a research process verb + preposition functioning as a cross reference to another section of the article.

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

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

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

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

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

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

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

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

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

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

A more typical spatial pattern involves technical biochemical processes including the classic expression 'in vivo'. 'Activity' for example usually takes place in 'organs':

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

'Concentrations' are only found in 'tissues' or 'tumours':

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

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

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

Proteins are followed by mutations, which are typically as we have said 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:

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

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

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

The third pattern we identify is the text referencing pattern, exemplified by the preposition's most frequent lexical left-collocate: 'shown in' (34 occurrences). The use of the present rather than past passive is noticeable in the following examples:

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

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

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

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

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

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

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

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

11.73 Results salient item 3: Did.

We discuss the role of 'did' together with the negative in the next section. Apart from this 'did' has two elliptical environments. The first after *but*:

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

The second is after comparisons of results and the item *than* (I have emphasised the elided empirical / biochemical process verb):

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

11.74 Results salient item 4: Not.

We have already noted the significant use of 'no' as a demonstrative presenting negative results in empirical or biochemical terms. Studying 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.

The most frequent right-collocate in of this expression is 'show', where the phraseology takes the form: (biochemical entity, usually living cells) did not show (biochemical process, usually treatment related):

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

There are expressions which go against this trend but they (as elsewhere in the corpus) are limited to formulations where the subjects are all data-related. Similarly, the very frequent right-collocate, 'differ', which emerges as a longer phraseology: [(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 the relational clause: *were not statistically significant*. This can be contrasted with affirmative relational clauses and uses of the verb 'show' when researchers tend to write that data is 'increased' or 'elevated'.

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

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

Drug therapies tend to 'induce' biochemical effects:

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

In a sense, when we have identified biochemical processes in the corpus, these have mostly appeared as technical verbs in very much the same distribution as nominalisations (c.f. induction of tumor necrosis factor). But we can also see that biochemical processes are expressed by more 'congruent' verbs, that is 'empirical process verbs', such as 'cause' and 'affect'. For example, despite its frequency, 'affect' is very specifically limited to the chemical process of binding:

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

Interestingly, in the passive the relationship is not symmetrical: the affecting medium tends to be a chemical or 'treatment':

accumulation was	not affected by	the treatment
relaxations were		nitro-L-arginine at any dose
reaction kinetics were		incorporation of cholesterol
excretion vomiting was		the presence of ...danorubicin
weight gain was		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
	further inhibition
	lysis
	any mortality
	tumorigenesis

Apart from biochemical or semi-biochemical processes, the negative in the results section is used to signal what the researchers didn't find. With 'was / were', we saw earlier that the passive in methods sections tends to be used with technical biochemical process verbs. Here a study of the negative reveals that the passive reverts to research process verbs and

that the passive, at least in negative voice, is usually modal in Results sections: (biochemical process) could not be (research process):

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

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

11.75 Results salient item 5: Had.

The role of relational processes such as 'is a' and 'have a' appears to be intimately linked with evaluation in this corpus. 'Had' is more restricted: and in the results subcorpus, 'had' serves to identify some degree of quantification rather than evaluation as in uses of *have / has* in Introductions. The subject often tends to be a biochemical subject:

Biochemical entity		Quantification	
mice	had a	decreased	number of formations
cells		different	correlation coefficient
animal tumours		greater	mean length
rat liver		higher	glucose count
patients		lower	frequency
protein		more pronounced	effect
infants		much lower	susceptibility
controls		normal	haryotype enzymes
LOH		significant	impact
subjects		smaller	body mass

This pattern has also been noted with 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		effect on the reduction of tumor size
treatment of narial cells		effect on weight gain
methanol control		effect on number of implantations
2 weeks experiments		effect on the factor X activator

One idiom that arises from this pattern is (tumour expression) had significant prognostic value:

Ta-T tumours	had significant prognostic value
tumor expression	
overexpression of p53	
The inhibitor	
The receptor antagonist ondansetron	

A similar pattern to one found in the abstract involves patients 'who had taken (drug X)', except that in the abstract the verb is usually 'had received'.

When 'had' is used as an auxiliary to express the passive perfect, we find a different pattern to the 'could not be + research process' pattern elsewhere in the results subcorpus. Here the verbs are clinical processes, mirroring the past passive with 'was /were' in the methods section:

electrode	had been	allocated
the film	had been	deposited [=left]
inspection of the electrode	had been	electropolymesmerised
tumour-bearing mice	had been	exposed to [x3 occurrences]
rats that	had been	treated to.

11.76 Results salient item 6: After.

After introduces clinical processes after some statement of time when the researchers are reporting quantitative results. Examples of typical phrases include:

expired	3 days after	injection
levels increased	10 hours after	completing infusion
levels were measured	4 hours after	PMEA administration
levels increased	4 hours after	drug administration
values were higher	24 hours after	completion of infusion

The most frequent phrase is *after treatment* (>50 occurrences). Apart from time periods, 'observed' is the most frequent left-collocate, and in some examples 'after' takes on its more frequent general language function of introducing time phrases:

the resistant phenotype	observed after	10 min. dilution time
the phenotype was		2 days cultivation
the resistance was		4 weeks of treatment
identical set of peptides was		induction by spiramycin
vermacin exposure was		administration of 8.5mg

'After' is also in the fixed expression 'after adjustment for' in order to briefly rephrase variables :

After adjustment for	other factors, we
	birth weight
	this additional variation
	tumor stage
	the same factors

11.77 Results salient item 7: There.

'There' has been seen in existential clauses in Abstracts and there signals past tense (and therefore current) evaluation of the change in the research articles' data shape. In Results, the most striking difference in use is the present tense and the preference for expressing quantitative rather than qualitative evaluation: the present tense is not used in other patterns with 'be'. Also unlike Abstracts, the most frequent pattern involves projections, where the main clause is generally a research process and the expression generally introduces evaluated empirical items:

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

'Found' is the most frequent of the above verbs as left-collocate, while the expression 'are considerable differences' is the most frequent right-collocate. There also appears to be a switch in tense with negative evidence or subjective statements or modality (but interestingly not without a modifier: *There was evidence of..*):

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

What phraseological principle can we postulate to explain why tense corresponds with syntactic choice in this way? One clue emerges in the phraseology of the verb group complex 'there appeared to be'. We have already noted that researchers tend not to use this

to 'hedge' but to signal a contrast (often preceded by 'Although'), and again we see this here:

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

We therefore have a set of grammatical choices that coincide with the negative 'There appeared to be no...' pattern:

- existential 'there'.
- modality.
- the use of the past tense.

What this may demonstrate is that the present tense pattern, with its thematicised research clause is a preferred way of presenting positive results, embedded within the modalised presentation of facts (we also note the number of non-hedged demonstrative references in the present tense / that-clause pattern: 'This shows that... This indicates that'). On the other hand, negative results may be presented as an aside or contrast with the main argument while the present tense indicates that an argument is to be taken forward.

11.78 Results salient item 8: The.

The significance of 'the' signals that textual reference to previously mentioned items presumably increases in later stages of the text, a discourse effect that correlates with increased lexical refocussing and rephrasing in later stages of writing. The definite article is obligatory in several collocational framework constructions, and so is a useful indicator. Among the more frequent frameworks, we identify the following categories:

Empirical framework:

(followed, increased, affected, reflected, mediated)
by the
(addition, method, end, presence, production)

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, implantation, removal)
of

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) at least the members of the bracketed items 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 our view that while we have observed wider patterns with prepositions elsewhere in the corpus, we have here the product of interaction between several smaller collocational relations.

11.79 Results salient item 9: when.

'When' is used to introduce subordinate clauses detailing a clinical process after a description of results and this usage suggests that some forms of subordination (especiually signalled by a conjunctive binder) increase in later stages of the research article. This is

much the same way that 'after' is used to introduce nominalisations of a clinical process. A frequent pattern involves ellipsis of subject and finite (Halliday's Mood) in reduced clauses especially with the verb 'compared':

Empirical measurement:		Clinical items:
were significantly reduced	when compared to	controls
yielded a 7 fold increase		the controls
showed superior effects		the same dose
yielded a 2-3 fold increase		the control group
there was a significant decrease		dose levels given

Although the distinction is not clear, a number of more biochemically oriented empirical results are modified by comparison 'with':

Empirical result:		Clinical item:
resulted in growth delay	when compared with	injection of saline
produced a significant effect		groups receiving no treatment
infusion was delayed		groups receiving no SCTT
therapy does reduce survival...		the same dose
prolongation of survival		those receiving...

When the main clause involves a research process, the subordinate clause includes the Mood, and since clinical processes are usually expressed by the passive this is the prevalent structure:

Empirical item:	Research process:	Clinical process:
loss of the film band	was observed when	films were photolysed
distinct redistribution		cells were treated
no activity		Hydrolam..was incubated
A 57% increase		H-7 was used
No significant effect		mice were treated

A similar phraseology involves the process 'obtained':

Empirical item:	Research process:	Clinical process:
show actual data points	obtained when	H-60 cells were exposed
The results	were obtained when	(X) was substituted
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

11.710 Results salient item 10: All.

'All' is a salient item in Results sections because researchers generalise across the totality their data. Of the more regular determiner patterns we find: 'in all cases' which precedes a statement of specific results (where the passive is not used with clinical processes):

In all cases	the medium was supplanted
	normal weight was regained
	the interval returned to baseline
	the relationship ... fell short
	nuclei had upfield shifts

The most frequent right-collocates are biochemical or related entities (groups, cell lines, patients, animals) and these tend to be introduced by the collocation 'all other':

all other	dose groups
	groups
	cell lines
	organs studied

The use of 'all' can be a good indicator of items that are on top of some hierarchy. In particular 'all other' functions as a deictic selective determiner signalling the kind of lexical refocussing we identified earlier. We also found that lexical rephrasing was higher in results sections, and this phrase may have a significant role to play in that. The following phrases are very typical of this, where 'all other' is not introduced by a preposition and is subject of the clause:

All other	dose groups of males were euthanized
	gross observations were checked
	microscopic findings were incidental
	microvessels showed no change
	regions remained the same in sensibility

However, if patterns are arranged by governing preposition, these reconfirm our earlier observation that prepositional phrases have a semantically limited set of constituents. Looking at them from the point of view of 'all' simply guarantees that we find items that are typically grouped together in the results:

tumors showed allelic loss	at all	information loci
One tumor had allele losses		3tm suppressor loci
we collected samples		time intervals
decreased		time points
IL2 secretion was inhibited		time points

11.8 Phraseology in PSC Discussion sections

Table 22: Discussion salient grammatical items from the Wordlist program

RANK	WORD	PSCDiscussion 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

11.81 Discussion salient item 1: That.

'That' is the most significant discussion salient item in the subcorpus. It is so atypical of the other rhetorical sections that it is listed in their word lists as among their least salient items, with the one interesting exception of Abstracts. In discussion sections, 'that' indicates the primary use of complement that-clauses that function as projections of research reports and facts (Halliday 1985:244). In terms of posture, 'that' introduces clauses that reformulate or evaluate results. That-clauses in the subcorpus can be divided into four patterns, in approximate order of frequency:

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 (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', has empirical measurement as subject, and the verb in the subordinate clause has some degree of modality or phase:

data suggests that
evidence suggests that
the model data suggest that
a number of observations suggest that
lack of ...activity suggests that

reactive oxygen would be important
simple sampling can be performed
endothelin receptors might play a role
MQ MT is unlikely to play a role in
patients should be monitored

On the other hand, 'indicate/s that' has deictic research process items as subjects and no modality in subordinate clauses:

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

Other verbs which share this non-modal phraseology are: 'show that, confirmed that' and 'demonstrated that'. Related to this structure, we find cleft clauses which are 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)
Another possibility (empirical)
One disadvantage (empirical)
The potential explanation (research)
The main conclusion (research)

The second pattern of complement clause 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 evaluation or negative results rather more 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		there are no differences in drug uptake
studies		the compound was not an inhibitor
		the parent compound was extensively metabolised
		active management was preferable

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

We conclude that	platinum orientation is not adequately represented
	CTL and NK cells together play an important role
	ifosamine is well tolerated
	MTT assay is suitable for assessing antiproliferative action
	this in vitro behaviour is meaningful

In our third phraseological type, extraposed it-clauses permit the researchers to omit the research process subject of the main clause, generally allowing for more modality in the main and embedded complement clauses. We have already suggested that this reveals a tendency for increased use of grammatical metaphor; in this case interpersonal 'modulation' metaphor. The most frequent left collocate for this pattern is 'possible', and its subordinate clause always has some modality (as does the NP complement: possibility that):

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

Instead of modality, negative polarity, or some negation of a previous sense occurs in the embedded clauses of 'it seems likely that':

it seems likely that	they missed the peak
	abnormal patterns affect [as opposed to normal ones]
	order and timing are not invariable
	cell counts were not carried in HMC100 p64
	... alterations did not reflect the PMN population

Extraposed clauses after the explicitly evaluative 'clear' do not have a preferred polarity but occur instead as themes of sentences introduced by adversative sentence adverbs:

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

The fourth pattern, embedded noun-clauses, appears to involve similar correspondences between verb and polarity or modality:

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

This expression forms a longer phraseological unit when it is introduced by clauses that express the modality of the proposition in terms of excluding it from or providing positive 'support' for introducing it into the research process:

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

A similar phraseology accompanies the phrase 'hypothesis that' where it is usually introduced by an empirical verb and the subordinate clause tends to have modality:

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

Another, 'evidence that' is introduced by evaluative negation (in the form of projecting clauses or negative epithets):

arguable whether any good — evidence that
there was no
there is no
The literature has failed to reveal
We found strong

The third most frequent NP complement is 'the fact that'. It appears to be intimately linked with negative results. In terms of semantic process and phraseology, the expression introduces embedded empirical-process clauses with negative polarity which then function as subjects of empirical remarks. In the first pattern, the negative result prompts speculation which stands as a result:

The fact that this enhancement does not occur in females
(implies that such oncogenes were not involved)

we cannot demonstrate this change
(suggests that AIN causes different effects)

the 150pp treated group was not killed earlier
(might be due to weakness in the dose monitor)

sequential accumulation of LOH was not observed
(might be due to early monitoring)

2 MCR lines did not show higher activity
(confirmed that these reagents were highly specific)

The expression '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 antibody columns 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
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) such as 'failure, ineffectiveness' generally 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. The idea that the 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': *'Insufficient sensitivity of our method [is the explanation of] the fact that we cannot demonstrate this degree'*. This suggests that research processes are not valid explanations and are not permitted by the phraseology and thus 'insufficient sensitivity' cannot be expressed as a negative result.

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,[sic]):

#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.

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
	(we) found that (+quantification)
it is possible that (+modal)	
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 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.

11.82 Discussion salient item 2: Be.

We have seen in the discussion of 'that' clauses that modality (in the evaluation of results) is a very salient feature of discussion sections. Since relational processes in the corpus tend to be involved in evaluation as well, it is not surprising that 'be' is the second most salient item. A corpus analysis of 'be' allows the analysis of the use of different modals. We describe the use of 'may' as a salient item later, but note that with 'be' it tends to be used for explanation rather than evaluation. When 'be' is introduced by 'can' the expression tends to be empirically oriented:

it	can be compared	with (x8 occurrences) to those (x4)

In the negative, the expression is uniquely used to express inclusion or exclusion in respect to the research model:

analysis	cannot be	excluded
range of interactants		completely excluded
ratio		ruled out

Whereas 'can' tends to be used in empirical expressions, 'could' tends to indicate either the researchers' ability to evaluate or explain a biochemical fact:

Biochemical process:		Empirical explanation / evaluation:
chemotherapy	could be	a potential benefit
chromatography		a promising candidate for
tumor expression		an appropriate target
This [inhibitor]		explained by two steps
This [overexpression]		explained as cellular

We also note an expression similar to the phraseology of 'due to the fact that': 'this discrepancy could be due to'. This overlaps with the overriding phraseology of 'must be' which also appears to explain:

Deictic biochemical/empirical process:		Empirical / biochemical explanation:
These results	must be	due to administration with
These results		due to reabsorption
This suggestion		due to enzymatic activity
The dispersion		due to seasonal variation
This variation		due to increased solvoyosis

This use of 'must' appears to differ from its exhortative or empathetic ('you must be tired') use in the general language (as described by Cobuild). Conversely, 'should' tends to be used to persuade or recommend, much as it is described in the dictionary, the difference being that the recommended actions tend to be research processes (or sometimes clinical or empirical processes, italicised):

	Research process:
should be	evaluated (x14 occurrences)
	investigated
	mentioned
	justified
	made with cations
	administered
	constantly under surveillance
They should undergo	further investigation

The exception to the Cobuild definition is 'it should be noted that' which corresponds to the Cobuild dictionary's description of 'must be noted that'. Here the projected subordinate clause expresses empirical quantification (mostly undirected):

It should be noted that	tumor cell lines are heterogenous
	others have found higher expression
	...tests have some degree of interdependence
	the degrees of inhibition... did not exceed 70%
	the decay does not take place in a concerted electron transfer

'Would be' tends on the other hand to introduce evaluation (with 'expected' being the most frequent right-collocate):

it	would not be	wise to allow plasma
the most likely source	would be	expected to return its reactivity
stretching modes	would be	sufficient
this localisation	would be	in agreement with
such a ...mechanism	would be	interesting to know

'Will' be also introduces evaluation, or an equivalent expression to 'expected' (in particular: required):

cytometric analysis	will be	required for different outcomes
samples		required to determine whether
this cohort		suitable
modulation of their kinase level		important for
their regulation level		necessary to estimate
tests		of limited value

A even more explicit distinction between evaluative and non-evaluative empirical processes emerges in examples of phase, 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 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'. This latter is in contrast with its non-evaluative active that-complement form (as noted above): '*We (have) found that'* + some degree of explanation (*we that the ester was more likely to be ionised*). The passive formulation'was found to be' involves more explicit evaluation by gradable adjectives:

(biochemical process X)	was found to be	considerably more potent
		more reliable
		safe
		the best strategy
		much higher

The 'evaluative' pattern is in contrast with that associated with expressions like the highly frequent '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		examined in order to establish
More.. cell tumors		studied in order to verify whether
These new strategies...		devised

A similar phraseology accompanies the phased (future) expression 'remains to be':

		(Research process)
its causal role	remains to be	determined
dependency		established
expressing different isomers		elucidated
whether this result in (x)		investigated
whether.. proteins would allow...		clarified

11.83 Discussion salient item 3: May.

We have already seen that 'may' and its relative 'might' are the main modals in subordinate clauses after expressions like 'it is possible that' and it is likely that'. In most of these expressions, modality was seen to correspond with explicit markers of evaluation. However, when the phraseology does not include 'it is possible that' or other types of subordination, the majority of the uses of 'may' appear to be true 'hedges', that is proposing an explanation but indicating to the discourse community that the researchers know it may not be true in all circumstances. We give two of the most frequent examples of this:

Empirical result:		Biochemical explanation:
ineffectiveness....	may be related to	sensitivity
efficiency of this line		crosstransformation
the more moderate effect		cell differentiation
derived cell lines		lower peak concentrations
this result		a bleeding tendency

lack of bioavailability	may be due to	error prone synthesis
deficiency in ..body weight		direct effects of replication
Another possibility		inherent differences in age
The fact we cannot demonstrate this charge		chance
Overexpression		disproportionate distribution

11.84 Discussion salient item 4: is.

'Is' is a salient item in introductions 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 process) to (research process)

In discussion sections, the patterns are less concentrated and 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)

Several expressions emerge as a new choice of wording in discussions, with less emphasis on thought processes and necessity, and more on affirmative evaluation:

It is	interesting to note that
	interesting that
	apparent that
	clear that
	most likely that

Extraposed rankshifted non-finite clauses are restricted to fewer patterns than observed in introductions:

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

An alternative attributive takes the form of an idiom: 'little is known about' which differs from the standard expression in introductions (X is known to):

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

Whereas in introductions, negative relational processes were concerned with negating the empirical relevance of biochemical processes (sensitivity is not detected,cholesterol is not applicable), here the tendency is to express negative evaluation of research processes:

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

Other attributive relational process patterns have specific phraseologies such as the pattern 'is consistent with' [research process is (empirical: consistent) with (research process)]:

This observation	is consistent with	all the results so far observed
This result		previous results
This		our findings
The level of protein found		findings obtained in ... models

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		in only in a minority of the tumor cells
peripherin		at high levels
protein		as micromoles
tumor size		by diameter

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

Biochemical process (disease related)		Empirical process
hypoglycaemia	is associated with	considerable increase in
The tumor mechanism		acquisition of t-cell properties
The MAC tumor		increased lactation
MOR phenotype		enhanced stability
Oncogene p185		internalization of bleeding

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

A reversed pattern emerges for 'is related to' which has as subject an empirical observation which is related to more specifically biochemically oriented items. This pattern shared by less frequent expressions ('is present in', and 'is responsible for'):

Empirical item		Biochemical / clinical process
risk	is related to	ethnicity
efficiency		stabilisation
the cause of toxicity		spasmodic polypeptides
presence of protein		expression of class III antigens
frequency in some tumor samples		the schedule of administration

11.85 Discussion salient item 5: Our.

Personal pronouns are infrequent in the corpus as a whole, and the appearance of 'our' is not surprising given that self-reference by the researchers ('we') also appears as a discussion salient item. 'Our' appears in a number of highly regular research process expressions, which we summarise as follows:

our	results	show/s that
	data	
	study	
	findings	
	studies	

Other verbs used in the phraseology are:

Our study	suggests	that
	indicates	
	demonstrates	

If the research term 'analysis' is used, no hedge or complement clause is introduced:

Our analysis	focused on a limited subset
	was based on immunohistochemical studies
	was based on four methods
	was to establish criteria for histology
	was to understand embedded tissue

Finally, an adjunct like 'clearly' is often used to emphasise the researchers' certainty if no 'hedging' verb (like *suggest*) is used

Our results	clearly indicate
	clearly demonstrate
	clearly show that
	strongly argue that

11.86 Discussion salient item 6: In.

The analysis of 'in' covers four of the six rhetorical sections in the corpus. In titles its left collocates were seen to be biochemical (metastases in, expression in, growth in) or empirical items (role of... in, change in). In abstracts, we noted a number of expressions involving empirical quantification (increase in, decrease in, reduction in, difference in). In results sections its use extended to quantification, a spatial use with biochemical entities and cross reference to other parts of the research articles. In discussion sections the tendency is

for empirical expressions of the shape of the data (the most frequent pattern) and causal relations (the second pattern). A third pattern involves research processes, and a fourth comprises several expressions where 'in' is involved in a phrasal discourse marker.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

11.87 Discussion salient item 7: Not.

Whereas in the results subcorpus, negative statements concerned causal relationships (*affect, cause, express*) and the general shape of the data (*increase, differ, was not different* etc.) the discussion sections express negative research observations. Again, unlike Abstracts negative data directions are not emphasised in Discussion sections, but 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 here it serves to report negative results:

we did not	detect	any	changes
we could	find		relationship
we did	observe		tumor development
we could	obtain		evidence of precursor
Early reports did	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.

Biochemical processes also appear in the framework 'not (biochemical process) with':

not	cross-react	with
	inserted	
	interfere	
	link	
	react	

The exception to this is the expression 'was not associated with', which includes the pattern: (quantitative empirical item) (be) not associated with (biochemical / research process):

amplification was	not associated with	...pathological characteristics
cell proliferation was		p53 overexpression
expression of cathepsin D is		response to endocrine theory
Patient's age was		nausea among our subjects
variations within the normal range are		the risk of developing ...disease

This shares a similar phraseology with the prepositional verb 'result in', where the second element is instead (biochemical / empirical process):

increasing data does	not result in	any further enhancement...
...native phosphate does		major conformational changes
estragon stimulation does		phosphorylation
Although the insertion mutation does		a form shift
substitution of the...backbone does		in large conformational changes

11.88 Discussion salient item 8: This.

In Chapter 10, we found that the deictic function of 'this' is a fundamental item in the process of reformulation in research articles. We found that deictic selective ('this' as head) and deictic lexical ('this' as premodifier of an anaphoric noun) are more frequent in discussion sections, and this is confirmed by concordance analysis. The most frequent pattern is as deictic selective:

This suggests that...
This may explain...
This might explain...
This is in agreement...
This is in contrast to...

In the deictic lexical pattern, any item modified by 'this' can be considered a reformulation, and the most frequent are anaphoric nouns reformulating previous text as research utterances (here they are listed by frequency):

Research reformulation by anaphoric utterance / cognition:

This result...
This finding...
This observation...
This model...[ambiguous: this may also be interpreted as a 'structure']
This hypothesis...

These are highly frequent, but are probably less varied than the terminological rephrasing of biochemical processes, of which the following are the most frequent:

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, we identify a series of patterns where the reformulations are superordinate items, used to relabel previous utterances. These patterns also clearly correspond with collocational frameworks, such as 'This (empirical reformulation of result) in (empirical item):

This	appearance	in	parental cells
	delay		PMN appearance
	difference		rate constant
	disparity		degree of suppression
	increase		metabolic rate

In the framework 'This...of' the pattern involves a superordinate empirical item which is itself measurable but does not constitute a result (as in the pattern above): 'This (empirical quantitative item) of (biochemical / empirical process/entity):

This	class	of	aromatic compounds
	comparison		ICso values
	dose		chemical...
	form		therapy
	group		tumours
	kind		analysis
	period		time
	range		concentrations
	type		damage

From our first list, we have omitted one high frequency item the is very frequently used to reformulate results, but is rather difficult to classify as either research or empirically oriented: this effect. As with 'empirical reformulation of result', effect can be seen to be empirical rather than research oriented because, as with all empirical processes, it labels observable and measurable phenomena (such as this motion, this reaction) and at the same time it could be construed as a researcher's interpolation or modelling of results (this tendency, this frequency). By reformulating results as an *effect* the researchers explain or comment on previous data-shape results 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.

The difference between the formulation 'this effect' above, and such research process expressions as 'This result' (as noted above, by far the the most frequent) suggests that 'this result' tends to introduce a new research direction that goes beyond comment on the previous clause or sentence:

#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 would seem that while most of these expressions are unique 'this effect' errs on the side of an empirical rather than research orientation. Yet again, we would claim that the semantics of a particular word are thrown into sharp relief by patterns of its use in this genre's phraseology.

11.89 Discussion salient item 9: We.

'We' is salient in discussion sections, and differs from its use in introductions sections in that it is used with more 'cognitive' research process verbs to do with postulations (conclude, believe, consider) whereas in the introduction it tends to be used with 'research

writing' processes to do with 'doing' (present, succeeded, compare). The difference is accentuated by the fact that, from our data on 'to' elsewhere, we find that action-oriented clauses are more typical of Introductions than propositional 'that' clauses in Abstracts. In discussions, '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...*). Of the 'cognitive process' patterns, which all occur in the present tense, we note the difference between the result-specific 'conclude' pattern which rephrases an empirical result and 'we believe that' which goes on to evaluate a negative result (italicised in sentences No.1):

Rephrasing
#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...

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.

11.810 Discussion salient item 10: Have.

We have seen that in Introductions 'has, have' are most often used with empirical items such as 'received, led to, attracted'. In the Discussions subcorpus, research processes are more emphasised. The majority of the research process uses of 'have' are described above, under 'we'. Passivised reports of research processes are the next most frequent use:

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

The next most frequent pattern involves active research processes:

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

As an attributive relational process, 'have' is used frequently to express evaluation, as seen elsewhere:

Biochemical /evaluative:		Evaluative:	Biochemical / Empirical process:
surviving cells	have	aberrant	morphology
the drug may		important	implications
the current assays may		limited	sensitivity
granisteron has been shown to		negligible	agonist
fragments have been reported to		superior	...localisation abilities

Part IV: Conclusion: The Discourse of Cancer

Chapter Twelve: Findings and Implications

This thesis has set out a possible corpus methodology for the analysis of genres, by establishing the context of situation of a specific type of scientific research article. It has demonstrated that by combining systemic grammar, discourse analysis and collocation analysis a comprehensive description of a specific genre is possible. The main methodological focus of the thesis has been the analysis of collocation, and the findings of this analysis are summarised here. This final chapter integrates the findings of the Data analysis and discusses the implications for the three research hypotheses set out in Chapter 6. The implications are then set in the wider context of scientific research writing and the chapter then goes on to discuss the effects of this in the wider community, including the popularisation of science and the discourse of cancer.

12.1 The reformulation hypothesis.

The reformulation hypothesis stated that new research ideas are created by the interaction of the textual processes of grammatical metaphor and discourse signalling (or posture). That is, the claim of a research article is built up by gradual changes in wording that correspond to processes such as nominalisation or lexical reformulation. We stated that the hypothesis is falsified if we could find no recurrent pattern of reformulation in posture or grammatical metaphor or no link between reformulation and the construction of scientific claims.

Reformulation is conceived in this thesis as a cohesive process where elements of discourse are reworded with a new discourse role. We followed the progress of a specific claim in ten research articles submitted by members of the cancer research group, firstly as the research claim is expressed by grammatical metaphor and secondly as the text progressively signals relations and encapsulates previous discourse in the research article. In the case of grammatical metaphor, we saw that new transitivity roles allow for increased terminological packaging. We also suggested that changes in transitivity roles (not necessarily congruent or metaphorical expressions in themselves) are significant points in the argumentation of scientific research articles.

In posture on the other hand, discourse itself is re-evaluated. Previous (and sometimes future) discourse is reformulated as new elements of terminological hierarchy or in terms of

the research activity. By explicit signalling of these relations or signalling of a change in argumentation, we have seen that the research article directs important elements of the discourse and to translate major research findings into a research model.

12.11 Reformulation and logogenetic history.

The analysis of logogenetic history reveals patterns of development of scientific claims. These patterns reveal that change in the expression of grammatical metaphor plays an important role in the textual construction of scientific claims. We find that change corresponds to rhetorical sections in research articles. Change involves the interaction of two mechanisms:

1) lexical reformulation of terminological and research utterances.
2) expression of empirical processes by relational or attributive clauses.

These findings are backed up by the observation of encapsulation, set out in the section below.

Specifically, we found that in half of the research article sample grammatical metaphor developed from congruent expressions. This was the direction postulated by Halliday and Martin (1993), a process that they claim represents a typical characteristic of written science. However, in the rest of the sample, we found that grammatical metaphor was cyclic (from congruent expression to grammatical metaphor and back to congruent expression) or occurred in reverse direction (from grammatical metaphor to congruent expression). We suggested that such a 'non-linear' logogenetic history of an expression is not just a linear process of construction, as Halliday states, but represents the rhetorical argumentation of the text. That is, if a typical logogenetic progression (congruent expression -> grammatical metaphor) represents what Halliday calls the genre of 'explanation', then alternative patterns such as the cyclical pattern represent significant shifts of argumentation. In Chapter 9 we saw that many of these shifts take place in different rhetorical sections. In some texts a cyclic or reversed pattern corresponded with the shifting of an expression to a higher research model (for example from *design concept* to *strategy* in JCPT10) or a shift from a general concept to a more specific term (e.g. from *consumption* to *oxidation* in BJ). This was particularly prevalent in results and discussion sections.

We suggested that grammatical metaphor includes the phenomenon of nominalisation (expression of verbal processes either as nominals or rankshifted clauses) as well as the process of lexical cohesion, especially Halliday and Hasan's (1976) 'general nouns'. Thus the re-expression of 'hydrolyse' as 'synthesis' is at once a nominalisation and a complex paraphrase (Hoey 1991). We found this kind of reformulation prevalent in the text sample, and found that it oriented around either terminological or research claims. As stated in section 9.3, Halliday's ideational metaphor involves reformulation within accepted hierarchies of scientific terminology. Thus the metaphor *synthesis* is itself reformulated more specifically as a *route* or more generally as a *mechanism*. The re-wordings we found throughout the text samples are clearly related the original expression, although their status as grammatical metaphor is questionable according to Halliday's, if not Martin's definition (Halliday 1985). Various conceptual directions of reformulation are afforded by these changes, whether they constitute grammatical metaphor or lexical paraphrase. As far as the discourse community is concerned, we argue that these terminological changes constitute small claims. This is demonstrated by the fact that grammatical metaphor allows for the insertion of evaluative terms in the new expression where before there are none (compare: *?this important synthesis* and the more likely *this important mechanism*). On the other hand, we find that reformulation often also involves a transition from terminological hierarchy to a broader research claim. By reformulating *mechanism* as a *strategy* in text TL, grammatical metaphor allows an empirical term to be worded as a research model: and we argue that this constitutes a stronger claim in the eyes of the discourse community.

Reformulation of terms is a contentious issue for scientists. Myers (1991) has demonstrated that discourse communities are defensive of previously held positions on terminology and that demotion of research claims by an editor often takes the form of terminological corrections. We find, however, that movement in terminology is visible in the building of longer blocks of nominals (as we saw in text CC). This confirms Halliday and Martin's (1993) assertion that grammatical metaphor plays an important role in the building of technical taxonomies. We have also observed the association of concepts (such as the link between *weight loss* and *glucose utilization* reformulated as *glucose consumption* in TPS) by relational clauses and empirical phrases (*are associated with*). Such associations typically correspond to changes in the direction of logogenesis. While some terms are reformulated as paraphrase or as research models, others remain constant. We suggest that grammatical metaphor is used most typically with unmarked biochemical terms that are 'assumed' in explanations or marked 'empirical' processes that are 'at stake' in the argumentation of the text. Thus biochemical entities tend to be quickly metaphorised in

argumentation of the text. Thus biochemical entities tend to be quickly metaphorised in introductory sections and remain so. On the other hand empirical processes (often realised by relational / attributive clauses) remain congruent. In certain marked expressions, however, we have seen that certain biochemical entities are untypically expressed 'congruently' (*flips, puckers* etc) and empirical processes are expressed metaphorically (*this relation, this association, the use of...*). We suggest that in such marked expressions, the concept is 'at stake' in that its status as fact has yet to be established by the text. This is borne out by our analysis of phraseology, which shows that empirical or research-oriented reformulations such as 'this effect' or 'this result' are subjects of attributive clauses which establish the empirical relevance of the previous (congruent) formulation (*this effect could be the result of...is even more pronounced, ...misleading, ...independent, this result may indicate that...eliminates the possibility that...is consistent with...*). Thus grammatical metaphor does not in itself make processes into facts, it enables expressions to modify stated facts.

So while we have identified grammatical metaphor as a fundamental linguistic mechanism of reformulation, we have not found that grammatical metaphor itself builds claims. Rather, new scientific facts rely on change within the text whether change of expression tends to be towards grammatical metaphor or not. We have already seen that there are few stable linguistic clues to the assignment of rhetorical 'moves' (Swales' 1981, 1990), although it may be that reformulation constitutes a strong clue to the staging of argumentation within broader rhetorical moves. However, we only have evidence of a correspondence between reformulation and main rhetorical sections. This would appear to support the reformulation hypothesis to the extent that rhetorical sections represent broader rhetorical moves necessary for the construction of claims.

To summarise, our analysis of ten sample texts suggests that logogenesis has at least an important role to play in the argumentation of the text, if not the textual creation of claims. Grammatical metaphor serves to aid explanation, to change the technical taxonomy and to promote empirical facts to research claims. We claim therefore that it is an important mechanism in the textual creation of new science.

12.12 Reformulation and posture.

In Chapter 10, we found that encapsulation of previous discourse is a useful model of reformulation. Posture concerns the sentence-to-sentence progression of discourse

signalling by deictic or conjunctive cohesion throughout the text. We argue that posture and grammatical metaphor interact: that metaphor is an important mechanism for encapsulating previous discourse. Our analysis of posture reveals patterns of encapsulation and prospection that correspond to rhetorical sections in research articles across the text sample. Explicit discourse signalling corresponds to the interactive plane of discourse (Sinclair 1981) and involves lexical reformulation of claims as terminology and research utterances (as seen above for grammatical metaphor). Lexical reformulation is thus a fundamental characteristic of encapsulation, which is in turn an important mechanism in formulating new science. In addition, the difference between different posture types may account for the variable textual autonomy or linear characteristics of certain rhetorical sections.

Specifically, in the ten pharmaceutical science texts, signalling tended to move from deictic refocussing of items (in abstracts, introductions and methods) to lexical rephrasing (in results and discussion sections). Methods sections are characterised by implicit (i.e. zero) signalling, while results and joint results / discussion sections typically display 'verbal echo' (typically sentences which share three or more paraphrases). Verbal echo in results sections refers back to methods sections and this indicates that data are reformulated as results and then as research models in discussion sections. Thus results and discussion sections are where information is promoted to a different status in the text. As with grammatical metaphor, lexical rephrasing indicates a new position of a previous technical item in the technical hierarchy or the status of a body of text in terms of the research paradigm (e.g. *these findings, this strategy*). We find that stretches of text characterised by rephrasing tend to be in section-final and discussion sections. This corresponds to Martin's (1993) 'reporting' genre where terms are reformulated as results and 'research utterances'. We have also seen that grammatical metaphor of modality (*it is possible that*) often corresponds to lexical rephrasing, and this supports the idea that grammatical metaphor allows the expression of explicit evaluation. Alternatively, stretches of text where deictic refocussing presents a theme in linear progression tend to be in section-initial and introduction sections. This corresponds with Martin's (1991) genre of 'explanation' where the characteristics of a single item are set out by attributive or relational clauses. In these sections evaluation is expressed more often by conjunctive signals. Thus introductions define and explain concepts, methods sections present raw implicit information, results and discussion sections report and promote the information firstly as evaluated findings and then as new research models.

The textual progression of discourse signals represents an important organising feature of

argumentation in research articles. The correspondence between types of posture and rhetorical section supports the reformulation hypothesis in that generic patterns of explanation and reporting occur alongside types of discourse signalling. This is an important step in accounting for how researchers present their claims. Whereas genre analysis has identified underlying rhetorical structures, the analysis of posture identifies short-range relationships that represent the writers' intervention in the text: signalling adversative relations and reassigning roles to previous stretches of discourse.

The fact that there may be no explicit signalling in certain sections does not indicate that scientific text is simply homogenous, an inert list of facts or instructions where there is no textual interdependency. Rather than seeing the research article as a linear text, we confirm our original hypothesis that certain areas of the text are autonomous (especially latter sections) and allow the reader to browse through the text in a non-linear way. We see later that phraseology may also have a role to play in the coherence of non-linear text. We suggest that experimental and methods sections can be consulted in non-linear fashion while the work of assigning coherence relations lies mostly with the reader. At the same time, introductions define terms and set out research hypotheses, and as a consequence cannot be treated 'indexically'. Results sections are more autonomous: we have already noted that they reformulate preliminary data set out in methods sections, whereas discussion sections restate results and reformulate them as research utterances. Abstracts turn out to be hybrid: they explain the preliminary ideas as in introductions and then set out research models as in discussion sections. Sinclair's model suggests that at any one point a sentence represents 'the state of the discourse' - but we interpret this to mean that certain sections may be more autonomous than others. Posture indicates the cohesive autonomy or interaction of the text, and throws into relief the relative differences between rhetorical sections.

Sinclair proposes that encapsulation of previous discourse leads the text on in a dynamic way. We would go on to argue that the author intervenes in the text in the case of explicit encapsulation - whether this involves conjunctive or deictic cohesive links. Where signalling does not take place, that is where the relation is (to the scientist) self-evident, the direction of argumentation in the text is presumably less relevant - in other words the direction of the discourse is autonomous, the complement of the interactive plane. Finally, while we have shown that signalling in the research article is heterogeneous, we have also seen that cyclic patterns of reformulation correspond to the interactive plane: that is deliberate manipulations of the scientific claim.

12.2 The phraseology hypothesis.

The phraseology hypothesis postulates that collocational patterns correspond to rhetorical sections of cancer research articles. While the reformulation hypothesis argues that phraseological patterns should identify systematic changes in the lexicogrammar of the research article. Phraseology reveals that conventional formulations of the discourse change within the text, and that the set of communicative goals and semantic concerns of the genre correspond to a delimited set of linguistic expressions. Our collocational analysis of salient grammatical items demonstrated collocational frameworks and collocational cascades correspond to consistent rhetorical functions in thew text (such as explanation, reformulation). In addition, grammatical items are the most consistent elements of sometimes long idiomatic phrases, and when they change, they largely entail consistent semantic changes in lexis. We argue that grammatical items can be seen not only as closed-class items, but as the fundamental elements of organisation in phraseological units.

The analysis of collocation in the PSC corpus confirms that collocational patterns of high frequency grammatical items vary consistently across rhetorical sections. Firstly, far from being equally distributed across the text, grammatical items have highly significant distributions according to rhetorical section. Secondly, collocational variation across rhetorical sections affects most areas of grammar and many areas of discourse in the corpus. Infinitive clauses of projection (introduced by 'to': *has been shown to...*) typically occur in introductions, while projection in abstracts and discussion sections is typically finite (*it has been shown that...*). This represents the metaphorical expression of modality. In addition, negative polarity in abstracts tends to refer to the quantification of results (*did not decrease significantly*), while in results sections negation refers to qualification of negative results (*did not result in*). Variation takes place in the construction of nominal groups (where 'of' is a significant item), in the signalling of negative results ('but'), in the encapsulation of immediately neighbouring discourse ('this'), in evaluation in relational clauses ('is'), in research or empirically oriented clause complexes ('that' or 'to') or passives ('been'), in the quantification or qualification of results ('in'), in the role of modality ('be') and interpersonal metaphor ('it'). Collocational analysis shows that a small set of semantic categories (research, empirical, biochemical and clinical processes) is a productive set of generalisations about scientific phraseology. A thematic classification of phraseological patterns into this kind of limited semantic system may be a characteristic of other specialised genres. This interrelationship confirms the phraseological principle set out

in the hypothesis: phraseology is a function of the communicative functions and collocational restraints of language.

One example of the phraseological principle is the interdependency of tense with communicative goals and the phraseology of certain verbs. We confirm Oster's (1981) and Malcolm's (1987) findings on tense: the analysis of 'was' and 'is' in the PSC shows that the present is used to present previous results and the general research paradigm, the past tense is used to express current methods and results. In other verbs we see that the difference in tense usage emerges in phraseology. The subject of past tense phrasal verb 'led to' is always a research-oriented item (*these observations*) while the subject of present tense 'leads to' is biochemical or empirical (*response to DMT damage*). Statements of given fact about biochemical items are likely to be present tense: statements of new research (observations) are likely to be past tense. However, this only extends to biochemical and empirical verbs: research process verbs (believe, observe, conclude) are generally present tense (including present perfect and present passive voice). Thus correspondences between global grammatical choices and lexical phraseology indicate that we have identified new facets of what is essentially a lexicogrammar (Halliday 1985).

We also find that differences in phraseology often indicate the differing semantic role of lexical items. This is best illustrated by collocational cascades: phraseological patterns that extend from one phraseological unit to another. For example, we found that in methods sections (but not elsewhere) the past passive phraseology *were + participle* involves mostly clinical or empirical process verbs. Further, the framework *were ...by X* 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 by NMR spectroscopy*. It may be that the longer the phraseology the more specific the semantic subcategories become. Hunston (personal communication) has suggested that such patterns are organised in steps which can be contrasted with phraseology which progresses in unrelated chains. Interestingly, in his analysis of semantic categories which emerge from large corpora, Barlow (forthcoming) has argued for similar cognitive categories as the fundamental organising units behind grammar.

It might have been expected that most of the data we have analysed in Chapter 11 would be difficult to characterise: grammatical items are not expected to display statistically significant collocational patterns. Instead, however, we found a surfeit of data, and every grammatical item displayed a rich range of collocational data, from relatively variable collocational

frameworks, to fixed expressions and idioms. To summarise the lexicogrammar in a few words would belie the complexity of the data. However, there are some general tendencies which serve to demonstrate the communicative functions of the rhetorical sections in the corpus. The phraseology of titles tends to form complex biochemical or empirically-oriented nominal groups with treatments as qualifiers. The phraseology of abstracts tends to form compaction (rankshifted clauses and conjunctive frameworks) and quantitative reporting of results. Introductions contain perhaps the longest stretches of phraseology, generally reformulating previous research or evaluating previous concepts. Methods sections contain a variety of fixed expressions and idioms, but their phraseology is principally concerned with the sequence of clinical procedures. The phraseology of results sections is predominately concerned the qualitative reporting and reformulation of negative data. Finally, discussion sections reveal a phraseology of overt evaluation and explanation of (negative) data reformulated as empirical rather than biochemical process.

The global analysis of phraseology also reveals the essential idomatic characteristics of the corpus. While some typical phrases emerged through the analysis of high frequency salient items, others revealed a combination of recurrent as well as idiosyncratic expressions (such as the use of 'Forefront' *Forefront in this role is tumor necrosis factor* ...(found in the introduction of Text JNCI)). Of the items we found, the following fixed and idiomatic expressions demonstrate the typical phraseology of the different rhetorical sections of the research article (variable elements are in parenthesis, and salient items in rhetorical sections are underlined):

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

Abstracts

(However) the mechanism of action of (Compound Y) was shown to (complex nominal and fixed expression of report)
there was a significant increase in toxicity (quantitative report)
It is concluded that (fixed expression of report)
propagation did not increase (quantitative report)
subjects who receive active management (fixed embedded clause)
both normal and tumor cells (framework with conjunction)

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 (fixed expression of evaluation)
(Compound X) is stable to the action of (Compound Y) (expression of established fact)
use of agents such as dismutase (reformulating 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 and idiosyncratic expression)

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 procedure)
(data set) calculated from the bootstrap samples 24h after exposure to (fixed expression of procedure)

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)

Discussion

data suggests that reactive oxygen would be important (modified report of results)
This results 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 suggests that (including encapsulation)
we have found that (report)

12.3 Phraseology and discourse.

Even the summary in section 12.2 above cannot do justice to the complexity and depth of data that phraseology reveals in a delimited corpus. Below, one possible representation of phraseological data has been formulated (figure 2). It is a summary of the collocational cascades in abstracts: that is, we attempt to link the most typical expressions on the basis of the phraseological analysis in Chapter 11. Phraseological expressions are linked by salient items, and sometimes by frequent lexical items that emerge as common elements between

phrases. The figure below represents a graphic summary of the ways the different patterns may fit together in an abstract. The links have been made only on the basis of the evidence we have presented above, except that in our analysis there is no immediate evidence for the position of phrases such as *In this study..are discussed*, although these correspond to their actual use in the corpus. What is represented is not a template where each element can be slotted in indiscriminately. It represents an implicit model of the most common phraseological choices available to the cancer researcher in the specific subgenre of the abstract, and these choices are limited by the topic and some sense of the preferred direction that the phrases may take as a longer string. In Chapter 11 we refer to collocations that join longer phraseological units as *collocational cascades*. In the figure we have used the same term to refer to the graphic summary of phraseology. As can be seen in the diagram, the semantic movement of most of the expressions is from *research orientation* to *methods* and from *report* to the *shape of the data* (i.e. the quantitative report of abstracts' phraseology as shown in Chapter 11). The cascade thus represents the typical linguistic expression of ideas and presents us with a model instantiation of the lexicogrammar of abstracts in cancer research.

The phraseology hypothesis also claims that change in phraseology may have a hand in the creation of new science. We claim that collocational cascades provide a baseline from which changes can be considered 'new science' in much the same way that Pavel considers the new collocation as 'terminology-in-the-making'. The claim is that the cascade represents the 'generic' part of the abstract for the reader: it is possible that any variations from the cascade will attract the attention of the reader and ultimately determine what changes the research paradigm. As Francis says:

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

The patterns we have identified in the analysis of the abstracts subcorpus are not accidental. We have seen in Chapter 5 that there is now 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 in the literature review sees the semantics of the word as textually distributed and syntax as intimately linked with lexical knowledge. Similarly, Fillmore, Kay and Connor (1988) write of phraseology in terms of:

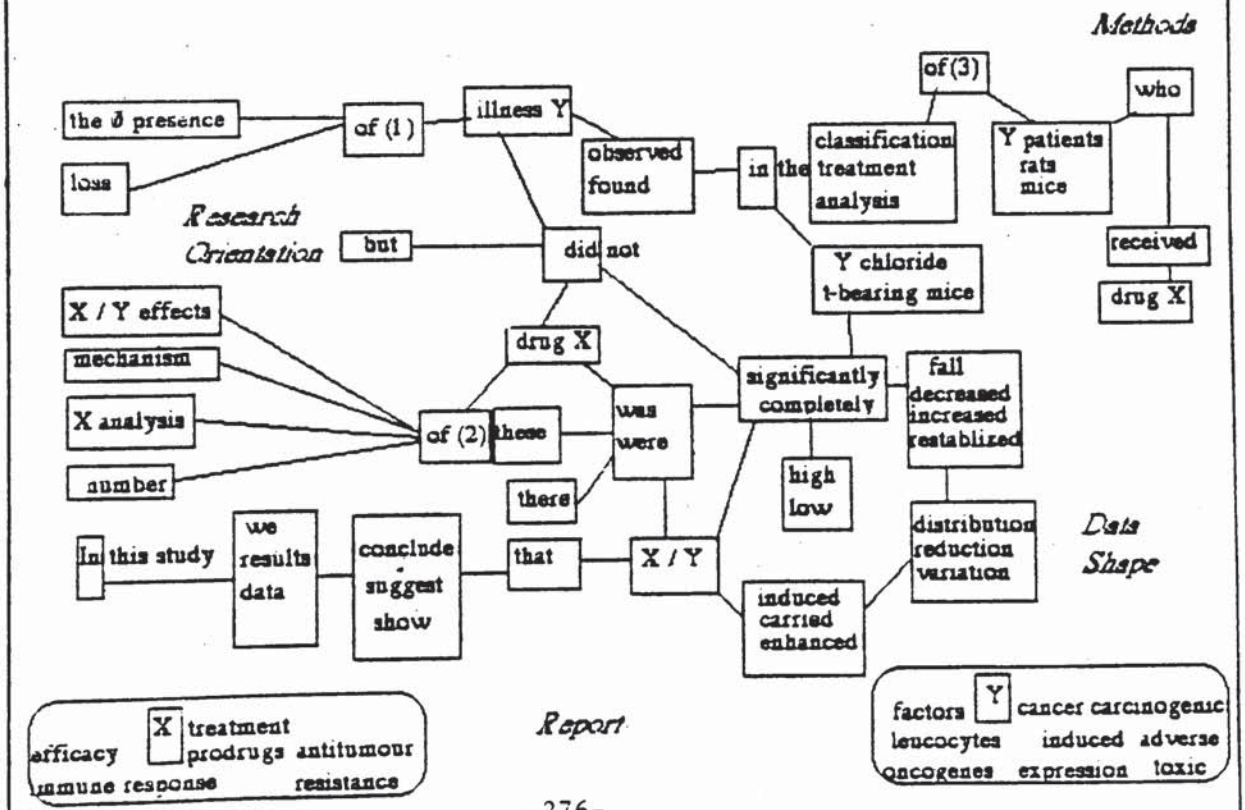
...phenomena larger than words, which are like words in that they have to be learned separately as individual facts about pieces of the language, but which also have grammatical structure [and] interact in important ways with the rest of the language. (1988:504)

In the specific context of cancer research articles, such instantial knowledge involves knowing which tense to use in expressing biochemical and research processes and even a subconscious knowledge of duality in the discipline introduced by *both* in abstracts. Instantial knowledge, represented in the formulation of phraseology, 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 progress from ideas to linguistic form:

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

Given this view, that meanings acquire their own wordings, we can therefore conceive of phraseology as the set of linguistic forms motivated by rhetorical aims and which further shape the discourse. It follows that the phraseological units we have identified are formulated by previous text and must have a role in the processing of the text as a whole. Clearly any changes in phraseology introduced by the author or any deviations from the collocational cascade must have consequences for the concepts throughout a running text, and we have demonstrated one aspect of this in the analysis of reformulation.

Figure (2): Collocational cascades in the cancer research abstract.



12.4 Integrating reformulation and phraseology.

What are the benefits of recurrent phraseological patterns in the business of scientific writing? We have hypothesised that phraseology is a key process that corresponds to conventional writing strategies in research articles. While rhetorical structure allows for accurate prediction on a broad scale, phraseological patterns may also be involved in what we have come to term the indexical function: the adoption of textual devices allowing for browsing and skimming though a text, as Nystrand suggests (1986). In our analysis of reformulation we have emphasised the notion of non-linearity by invoking 'rhetorical convention'. In their studies of signalling and use of rhetorical structure, Swales (1981), Nwogu (1989) and Sharp (1989) had found that predictable elements of rhetorical structure and visual format help readers to identify where to jump to, to guess the content of conventional areas of the texts. But while such analysis helps describe the linear reading of texts, it doesn't account for 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 Hoey's work (1991). So in addition to powerful tools of models of reformulation, rhetorical structure and format, it is worth considering whether grammatical parallelism, conventionalised phrases and cohesive networks might also be used in these texts to complement their non-linearity of use.

The systemic view of language sees all systems of language as ultimately connected, and our contention is that the phenomena we have identified fit on a cline. At one end there are 'short range' textual phenomena of formulations, represented in this thesis by terminology and collocation. At the other end there is a model of discourse analysis which embodies 'long range' issues of textual cohesion, rhetorical moves and ultimately the language system of a discourse community. Collocation and reformulation appear to fit somewhere between the two extremes. The cline does not represent the 'amount of context' required to provide evidence for these phenomena: a coherent understanding of short-range phraseology may require just as much specific contextual knowledge as long-range phraseology. The difference is largely co-textual: collocations can be observed locally, while reformulation can only be observed as it operates throughout the running text. However, we argue that collocation can only be seen to function in terms of reformulation and *vice versa*. Phraseology thus accounts for change in the language system by a process of continual reformulation by mechanisms such as grammatical metaphor and discourse signalling.

12.5 The adaptive science hypothesis.

The adaptive science hypothesis postulates that cancer research adapts linguistic processes to create its own unique form of language. We argue that science writing (in particular the research article) is the only means by which scientific change can be enacted, and that while this linguistic form is topically and rhetorically unique (that is it is restricted in use to a specialist discourse community), it is still part of the general language system. The argument is that there must be mechanisms of change within the language system that are adapted specifically to the purposes of science and that are key to the specific role of innovating new terminology and elevating concepts from data to new research models.

In the analysis of the context of cancer research writing in Chapter 7, we saw that cancer is a diffuse concept rather than an entity or a terminological fact. We argue that a researcher's perspective of cancer will determine how he or she envisages and presents his or her work to the outside world. We also saw in the survey that scientific activity resides just as much in the production and use of texts as it does in laboratories. In the indexical function, research articles are treated as repositories of instructions; in one researcher's words, research articles are 'cook books'. The indexical and referential use of the research article demonstrates that it is an integral part of laboratory activity. So similarly, the signalling of successive research claims, the taxonomising effect of grammatical metaphor and the changes in wording seen in the language of these texts also initiate action by the reader, even if this is conceptual: they augment the readers' knowledge and teach new phraseology. At any point where the reader acts upon the wording of a research article, he or she can be said to be engaged in scientific activity. Thus by postulating the adaptive science hypothesis we argue that science, embodied in complex concepts such as cancer, is language-like: science is not transmitted via language because it is already a form of language. We apologise for an analogy with genetics here: we conceive of science as a set of instructions couched in a discrete combinatorial system (language or DNA). The instructions or the codes in which they are set do not constitute scientific activity: it is the enactment of the code which constitutes science. Evoking natural selection here may appear far-fetched, but the analogy is striking: successful scientific ideas are only replicated by attracting readers rather than being intrinsically useful. The burden of usefulness is ultimately placed on the reader, and the point of interaction between the reader and text is enabled by reformulation and phraseology.

The adaptive science hypothesis is formulated to complement the constructivist view of

scientific writing, proposed in a number of language-related fields by Knorr-Cetina (1983), Swales (1990), Halliday (1993), Myers (1990) and others. Constructivism argues that science is negotiated by the dynamics of the discourse community. Within the discourse community, claims (which we have seen in the survey in Chapter 7 are equivalent to relative novelty of scientific knowledge) are diminished by editors in terms of more acceptable terminology, as we note above. What the constructivist approach does not identify are the exact linguistic mechanisms of change in single texts; that is, how science writing establishes information, builds argumentation on the basis of the original information and how this information is changed by the text. So, while rhetoricians have identified the characteristics of successful texts, we have identified some mechanisms that allow the correct interpretation and use of such texts. Whether the text is useful or successful or not is thus not our primary concern: we are interested in how science is reconstructed within the text. Similarly, geneticists analyse how the 63-letter code of DNA is used to construct complex context-sensitive compounds: they are not interested in whether those compounds are beneficial to the organism. So with apologies to Richard Dawkins, we are conducting developmental as opposed to evolutionary linguistics.

The processes of grammatical metaphor (Halliday and Martin 1993, Martin 1991) and terminology-in-the-making (Pavel 1993) have been proposed as possible mechanisms for change in science writing, and we have argued above that these mechanisms are fundamental to building scientific knowledge as text. We have also suggested that devices such as reformulation may be exploited in scientific articles to allow for the indexical use of texts. In an ethnographic survey of the context of scientific writing, we have already seen that non-linear uses of text are enabled by specific format and terminological devices which enable cancer-specific information to be encoded. It is not difficult to demonstrate that these are unique textual devices. The difficulty lies in proving that normal processes of rhetoric are adopted and changed to suit genre-specific purposes, and that the resultant tailor-made processes are unique. Since it is beyond the scope of this thesis to make claims about rhetoric, our aim is to extend the adaptive science hypothesis only to the smaller-scale area of linguistic devices. 'Uniqueness' does not imply the kind of 'language for special purposes' concept used by certain terminologists. Terminologists such as Sager et al. (1990) claim that some linguistic devices are 'special' or unique in the specific scientific activity. We are not arguing that certain linguistic resources are unique, instead the hypothesis is that it is distinctive use of resources rather than the nature of the resources themselves that constitutes English for Specific Purposes.

12.6 Popularisation and the discourse of cancer.

> 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)

We have so far stated that the cancer research article functions by reformulating assumed concepts and creating new phraseology. In a broader hypothesis, we have suggested that science is reproduced as language using the mechanisms of reformulation and phraseology. The first aim of this thesis was to discover how language functions in highly specialised circumstances, although now that we have argued that science is inseparable from language, it is perhaps appropriate to briefly analyse the interaction between the esoteric world of research article writing and the genres and other activities which interact with the research article. Halliday and Martin have argued that the influence of scientific discourse is pervasive in society, especially in 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 English.

While other forms of discourse may be equally as pervasive (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. In Chapters 2 and 3 we saw that research articles compete with review articles, experimental articles, accelerated communications, 'popular' science articles (in Nature etc.), indexing abstracts, and other genres for the attention of the immediate research community, but we also note the important role of the grey literature, of grant proposals and the reports of the research funding councils and the press agencies of the major cancer charities. The local and national press are informed at regular intervals of research 'breakthroughs' and the cure for cancer story on television is carefully timed at monthly intervals to emerge from different Universities and research institutes (*MT*, personal communication). Typical of the upbeat, fund-raising character of 'intermediary' publications is the Association for International Cancer Research's newsletter *Progress*, which presents popularised explanations of the work of several research groups with regular rubrics for contributors such as 'The meaning of cancer'. These articles are closest to the newspaper articles we find reporting research from Aston's cancer research group. As noted in Chapter 7, the group has had a number of 'breakthroughs' relating to *MT's* findings reported in texts JNCI, BJ and TPS. In future research we hope to report on the types of reformulation that take place from highly specialised texts such as JNCI to an articles such as the Daily Telegraph's 'Cancer discovery by farmer scientist'. Initial

findings suggest that much of the science is lost to discussion of personalities involved (most especially the role of charities and fund-raisers and the lifestyle of the senior researchers), the national or local role of the research group (depending on the national or local scope of the reporting newspaper) and finally issues not presented in the original research such as the meaning and fear of cancer. Those parts of the articles which do report the actual research findings emphasise the novel approach to what is presented as an old problem and the discovery of a new substance (even though in MT's case the characterisation of a process *cachexia* is the real discovery). The result reported in BJ is effectively formulated as:

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 reporting this finding in the local and national press, the first sentence of the Independent suffices to show the processes of reformulation that 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)

We have emphasised the expressions and phraseology which are reminiscent of our corpus: the degree of compaction is high (the use of 'of', post-modifying reduced relative clauses) as is grammatical metaphor (one passive construction and nominals: treatment, new evidence, weight loss). However, this is perhaps not surprising since journalists themselves use the Charities' own press releases to digest the findings. The consequences of this are not yet clear. But what we have seen in research articles alone would seem to suggest that stereotypical issues of scientific writing such as nominalisation, passivisation and general complexity of expression cannot be simply seen as characterisations of the scientific language or research article register, but must be seen as important processes that govern our way of writing these thoughts.

12.7 Further research

One application of the analysis of collocation in genre analysis may be that phraseological patterns are acquired piecemeal by the slow processes of re-editing and re-reading that apprenticeship in the discourse community requires (Myers 1990). The whole phrase acquisition of language is a process that has been proposed by several researchers, such as

Pawley and Syder (1983), Peters (1983) and Widdowson (1989). It may be possible, for example, for genre analysis or ESP to make some use of 'collocational cascades' as short cuts in order to save time when teaching English for reading and producing research articles as Johns and King (1993) have suggested. The 'lexical syllabus' has become a focus of debate in second-language teaching (Willis 1990) and the work of Cobuild has led to a number of phraseologically based teaching materials (Willis 1993). In addition, the slower, immersed approach to acquiring phraseology may be a useful analytical tool, not only in monitoring the linguistic progress of apprentice writers, but also in analysing how texts are edited, how coherence develops chronologically throughout a text and how phraseology evolves over time, just as Atkinson has demonstrated with rhetorical structure (1992). Earlier, we suggested a genetic metaphor for language and the natural selection of scientific ideas and proposed the term 'developmental linguistics' as a cover term for the diverse areas that accompany and include 'applied linguistics' but which essentially attempt to marry pragmatics with linguistics. We have already proposed a phraseological view of logogenesis, and would like to suggest that future work be applied to ontological development (especially genre acquisition in the individual) and phylogenic development (evolution of genre in the discourse system), although these belong to another thesis.

BIBLIOGRAPHY

AARTS J. 1991 " Intuition-based and observation -based grammars." in K. Aijmer and B. Altenberg 1991 :44-62

AARTS J. 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

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

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. , HOLMES-HIGGINS P., ROGERS M., and THOMAS P. 1989 "The Translators' workbench project." in Translating and the computer II Proceedings of ASLIB 16-17 Nov. 1989 London.

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

AIJMER K. 1986 "Discourse variation and hedging." in J. Aarts and W. Meijs 1986: 1-18

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

ALLEN P. , McNEAL M. and KVAK D. 1992 "Perhaps the lexicon is coded on a function of word frequency?" in Journal of Memory and Language Vol. 31: 826-844

AL-SHEBAB O. 1989 "Organisational and textual structuring of radio news discourse in English and Arabic" Unpublished PhD. Thesis, Language Studies Unit, Aston University

ALTENBERG B. "Amplifier collocations in spoken English." in S. Johansson and A.B. Strenström (eds.) :127-147

ALTERMAN R. 1990 "Some computational experiments in summarization". in Discourse Processes Vol 13 :143-174

ALTERMAN R and BOOKMAN L.A. 1992"Reasoning about a semantic memory encoding of the connectivity of events." in Cognitive Science Journal 16/2

ANDOR J. 1989 "Strategies, tactics and realistic models of text analysis." in W. Heydrich et al. 1989 :28-36

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

AUGER C.P. (ed.) 1975 Use of Reports Literature London: Butterworth

AUSTIN J.L. 1962 * 1975 (eds. Urmson J.O. and Sbisà M) How to do things with Words London: Oxford University Press

BAAKES K. 1992 "A communicative approach to teaching terminology in ESP" in Fachsprache Vol. 14/1-2 :23-40

BAHNS J. 1993 "Lexical collocation: a contrastive view." in English Language Teaching Journal Vol. 47/1 :56-63

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

BANKS D. 1994 'Clause organization in the scientific journal article' ALSED-LSP Newsletter Vol. 17/2 :4-16

BARBER C.L. 1962 "Some measurable characteristics of modern scientific prose." in Almquist and Wikwell (eds.) Contributions to English Syntax and Philology :21-43

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

BARRY A. 1993 "A computational analysis of school letters to parents. A passport for entry into a limited discourse community." Unpublished MSc. Thesis, Language Studies Unit, Aston University.

BARTHES R. 1966 Mythologies Paris: Seuil

BARTLETT R. 1932 Remembering Cambridge University Press

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

BEAUFRÈRE-BERTHEUX C. 1994 'De l'anglais médical.' ALSED-LSP Newsletter Vol. 17/2 :17-23

BÉJOINT H. 1988 "Scientific and technical words in general dictionaries." in International Journal of Lexicography Vol. 1/4 :354-368

BENGT H. and ALTENBERG B. 1990 "Phraseology of spoken English: Presentation of a project". in Aarts J. and W. Meijs (eds) 1990 :1-26

BENSON M. 1989 "The collocational dictionary and the advanced learner." in M.L. Tickoo (ed.) Learner's Dictionaries: State of the Art Singapore: SEAMEO Regional Language Centre :84-93

BENSON. M. 1990 "Collocations and general purpose dictionaries." in International Journal of Lexicography Vol. 3/1 :23-25

BENSON. M., BENSON., E. and ILSON R. 1986 The Lexicographic Description of English London: John Benjamins

BENVENISTE E. 1966 Problèmes de Linguistique Générale Paris: Gallimard

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 M. 1977 Introduction to Systemic Linguistics London: Batsford

BERRY-ROGGHE G. 1970 "Collocations: Their computation and semantic significance." Unpublished Ph.D thesis UMIST, Manchester

BERRY-ROGHE G. 1973 "The computation of collocations and their relevance in lexical studies." in A.J. Aitken and R.W. Bailey (eds.) The Computer and Literary Studies Edinburgh: Edinburgh University Press: :103-112

BIBER D. 1986 Variation across Speech and Writing Cambridge: Cambridge University Press

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

BIBER D. and FINEGAN E. 1989 "Styles of stance in English: Lexical and grammatical markers of evidentiality and affect." in Text Vol. 9/1 :93-124

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.) 1991a Sociolinguistic Perspectives on Register Oxford: Oxford University Press

BIBER D. and FINEGAN E. 1991b "On the exploitation of computerized corpora in variation studies." in K. Aijmer and B. Altenberg 1991 :204-220

BLACKWELL S. 1987 "Problems in the automatic parsing of idioms." In R. Garside et al. (eds) Syntax versus orthography :110-119

BLANTON 1982 "The pragmatic structure of rhetorical maturity in the sciences." in W. Frawley (ed.) 1982: 128-143

BOBROW D. and COLLINS A. (eds.) Representation and Understanding New York: Academic Press

BODEN M.A. (ed) 1990 The Philosophy of Artificial Intelligence Oxford University Press.

BODEN M.A. 1990 "Escaping the Chinese room." in Boden (ed.) 1990 :98-103

BOGURAEV B. and BRISCOE T. 1987 "Large lexicons for natural language processsing: Utilising the grammar coding system of LDOCE."

BORKO H. and BERNIER C.L. 1975 Abstracting concepts and methods New York Academic Press

BORKO H. and CHATMAN S. 1963 "Criteria for acceptable abstracts: a survey of abstractors' instructions." in American Documentation Vol. 14 :175-184
BRANSFORD J.D. and FRANKS J.J. 1971 "The abstraction of linguistic ideas." in Cognitive Psychology Vol.2 :331-350

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

BROEK P.V.D. and TRABASSO T. 1986 "Causal networks versus goal-hierarchies in summarising text." in Discourse Processes Vol. 9/1 :1-16

BROWN G. and YULE G. 1983 Discourse Analysis Cambridge University Press

BROWN P.F., DESOUZA P.V., MERCER R.L., PIETRA J.D. and LAI J. 1992 "Class-based n-gram models of natural language." in Computational Linguistics Vol. 18/4 :467-479

BRUCE N.J. 1983 "Rhetorical constraints on information structure in medical research writing." Paper presented at the ESP in the Arab World Conference, University of Aston, UK. August 1983

BUSCH G. 1992 "Search and Retrieval." in BYTE, June: 274-282 Bix Publishers

BUTLER C. 1985b Statistics in Linguistics Oxford: Basil Blackwell

BUTLER C. 1985a Computers 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

BURNARD L. 1992 "Tools and techniques for computer-aided text processing." in C. Butler (ed.) :1-28

BURROWS J.F. 1992 "Not unless you ask nicely: the interpretative nexus between analysis and information." in Literary and Linguistic Computing Vol. 7/2 :99-109

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

CARTER R. 1988 "Front pages: lexis, style and newspaper reports." in M. Ghadessy (ed.) :8-16

CARTER R. 1992 "Lexis." in Research in English Language Teaching Vol. 2/1 :85-99

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

CHARGAFF E. 1986 "How scientific papers are written." in Fachsprache Vol. 8 :106-110

CHARNIAK E. and WILKS Y.1976 Computational Semantics Fundamental Studies in Computer Science No.4, Amsterdam: North Holland Publishing Company

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

CHURCHLAND P.M. 1991 "Some reductive strategies in cognitive neurobiology." in Boden (ed.) 1991 :334-367

CLEAR J. "Overview of the role of computing in Cobuild." in J.McH.Sinclair (ed.) 1987 :41-61

CLEAR J. "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

CORSON D. 1985 The Lexical Bar London: Pergamon Press

COTTRELL G.W. 1989 A Connectionist Approach to Word-Sense Disambiguation London: Pitman

COULTHARD M. (ed.) 1986 Talking about text Discourse Analysis Monologue. No. 13, Birmingham: English Language Research, University of Birmingham.

COULTHARD M. (ed.) 1992 Advances in Spoken Discourse Analysis London: Routledge

COWIE A.P. 1981 "The treatment of collocations and idioms in learners' dictionaries." in Applied Linguistics Vol. 2/3 :223-235

COWIE A.P. 1989 "Multiword lexical units and communicative language teaching." MS For the International Colloquium on Vocabulary and Artificial Intelligence, Université Lumière, Lyon, 1989.

CRAVEN T.C. 1965 "Sentence dependency structures in abstracts." in Library and Information Science Research Vol. 10 :401-11

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

CROOKES G. 1986 "Towards a validated analysis of scientific text structure." in Applied Linguistics Vol. 7/1: 57-70

CROSLAND A.T. 1975 "The concordancer and the study of the novel." ALLC Bulletin Vol.3 :190-196

CRYSTAL D. 1991 "Stylistic Profiling." in K. Aijmer and B. Altenberg (eds) 1991 :221-238

CRUSE D.A. 1986 Lexical Semantics Cambridge University Press

CRYSTAL D. 1991 Stylistic profiling. in K. Aijmer and B Altenberg (eds) 1991 :221-238

DAHL Ö. and FRARUD K. (eds.) 1988 Papers from the First Nordic Conference on Text Comprehension in Man and Machine Proceedings, Oct. 27-28 1988 Institite of Linguistics: University of Stockholm.

DAVID J. and MARTIN R. (eds.) 1977 Etudes de Statistique linguistique Metz: Centre d'analyse syntactique de l'université de Metz

DAWKINS R. 1986 The Blind Watchmaker Harlow: Longman Scientific and Technical

DE BEAUGRANDE R. "Psychology and composition: Past, Present and Future." in M. Nystrand (ed) 1982: 211-267

DE BEAUGRANDE R. and DRESSLER W. 1981 Introduction to Text Linguistics London: Longman

DELISLE J. 1984 L'Analyse du Discours comme Méthode de Traduction Ottowa: Editions de l'Université d'Ottowa

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

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

EDGE J. 1993 "The dance of Shiva and the linguistics of relativity." in Applied Linguistics Vol. 14/1 :43-55

ENDRES-NIIGGEMEYER B. 1985 "Referierregln und Referate- Abstracting als regelsgesteuerter Textverarbeitungsprozeß." in Nachtrichten für Dokumentaristen Vol. 36/1 :38-50

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

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

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

FIDEL R. 1986 "Writing abstracts for free-text searching." in Journal of Documentation Vol. 42/1 : 11-21

FILLENBAUM S. and RAPOPORT A. 1971 Structures in the Subjective Lexicon New York: Academic Press

FILLMORE C.J. 1968 "The case for case." in E. Bach and R.T. Harms (eds.) Universals in Linguistic Theory New York: Holt, Rinehart and Winston: 1-88

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. 1957 Papers in Linguistics 1934-1951 Oxford: Oxford University Press

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

FOX G. 1987 "The case for examples." in J.McH. Sinclair (ed.) 1987 :137-149

FOX G. 1993 "A comparison of 'policespeak' and 'normalspeak' : a preliminary study." in J. McH. Sinclair et al. (eds.) 1993 :184-195

FOXLEY G. and GWEI M. 1989 "Synonymy and contextual disambiguation of words." in International Journal of Lexicography Vol. 2/2 :111-134

FOZ C. and I VAZQUEZ 1993 "The persuasive function of lexical cohesion in English: a pragmatic approach to business reports." unpublished MS presented at the 5th International Systemic Workshop on corpus-based studies, Universidad complutense de Madrid, 26-29 July 1993

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

FRANCIS W.N. 1992 "Language Corpora B.C." :17-32 in J. Svartvik (ed) 1992a

FRANCIS W.N. and KUC'ERA H. 1982 Frequency Analysis of English Usage: Lexicon and Grammar Boston: Houghton Mifflin

FRAWLEY W. (ed.) 1982 Linguistics and Literacy London: Plenum Press

FROHMAN B. 1990 "Rules of Indexing: A critique of mentalism in information retrieval theory." in Journal of Documentation Vol. 46/2 :81-101

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.

GALE W.A., CHURCH K.W. and YAROWSKY S. 1993 "A method for disambiguating word sense in a large corpus." in Computers and the Humanities Vol.26 :415-439

GARROD S. 1986 "Language comprehension in context: A psychological perspective." in Applied Linguistics Vol. 7/3 :226-238

GARSIDE R., LEECH G. and SAMPSON G. (eds.) 1987 The Computational Analysis of English: a corpus-based approach London: Longman

GELIPITHIS P.A.M. 1988 "Survey of the theories of meaning." in Cognitive Systems Vol. 2/2 :141-162

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

GHADESSY M. (ed.) 1988 Registers of Written English: Situational Factors and Linguistic Features London: Frances Pinter

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

GIORA R. 1990 "On the so-called evaluative material in informative text." in Text 10/4 :299-319

GLÄSER R. 1991 "The LSP genre abstract - revisited." in ALSED - Newsletter Vol. 13/4 :3-11

GLÄSER R. "A multi-level model for a typology of LSP genres." in Fachsprache Vol. 15/1-2: 18-26

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

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

GNUTZMANN L. and OLDENBERG H. 1992 "Contrastive text linguistics in LSP research: Theoretical considerations and some preliminary findings." in Schneider (ed.) :103-136

GODLY 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

GRAETZ N 1985 "Teaching EFL students to extract structural information from abstracts." :225-335 in J.M. Kline and A.K. Pugh (eds.) Reading for Professional Purposes: Methods and Materials in Teaching Languages

GREENBAUM S. 1991 "The development of the International Corpus of English." in K. Aijmer and B. Altenberg (eds.) 1991 :83-91

GRICE H.P. 1975 "Logic and Conversation" in P. Cole and J.Morgan (eds.) Syntax and Semantics III New York: Academic Press

GRIMES J.E. 1975 The Thread of Discourse The Hague: Mouton

GRISHMAN R. 1986 Computational Linguistics Cambridge: Cambridge University Press

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

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

HALLIDAY M.A.K. 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 Harmondsworth: 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 probalilistic 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 Svartvik (ed.) 1992a :61-77

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

HALLIDAY M.A.K and MATTHIESSON C. 1993 Ms: Construing experience through meaning: a language-based approach to cognition.

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

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

HAUSSMANN F.J. 1989 "Le dictionnaire des collocations." in F.J. Haussmann, O. Reichmann, H. E. Wiegand and L. Zgusta (eds) Wörterbücher, Dictionaries, Dictionnaires Volume 1, Berlin: De Gruyter

HAYES J.R. and FLOWER L.S. 1980 "The Dynamics of Composing." in L.W. Gregg and E.R. Steinberg (eds.) Cognitive Processes in Writing New Jersey, Lawrence Erlbaum Associates

HAZADIAH M.D. 1993 "Topic as a dynamic element in spoken discourse." in Baker et al. (eds.) 1993 :55-74

HEIDEGGER M. 1966 Discourse on Thinking London: Torch, Harper and Row

HEURING V.P. 1985 "The automatic analysis of fast lexical analysers." Unpublished PhD Thesis, University of Colorado: Dept of Electrical and Computational Engineering.

HEYDRICH W., NEUBAUER F., PETÖFI J.S. and SÖZER E. A. 1989 Connexity and Coherence: Analysis of Text and Discourse Berlin: de Gruyter

HINTON G., McCLELLAND J., and RUMELHART D. 1986 "Distributed Representations." in M. Boden (ed.): 248-280

HIRAYAMA K. 1964 "Length of abstract and amount of information." in Journal of Chemical Information Vol. 4 :9-11

HIRSCHMAN P. GRISHMAN R and SAGER N. 1976 'From text to structural information processing of medical reports'. Proceedings of National Computer Conference. Associated APS Press: Mountvale New Jersey :267-275

HOEY M. 1979 Signalling in Discourse Birmingham: University of Birmingham English Language Research monographs No. 6

HOEY M. 1983a On the Surface of Discourse London: Allen and Unwin

HOEY M. 1983b "Three metaphors for examining the semantic organization of monologue." in Analysis: Quaderni di Anglistica Vol. 1/1 :27-54

HOEY M. 1986 "Overlapping patterns of discourse organization and their implications for clause relational analysis of problem-solution texts". In C. R. Cooper and S. Greenbaum (eds.) Studying Writing: Linguistic Approaches, California: Sage Publications

HOEY M. 1988 "The clustering of lexical cohesion in non-narrative text." in Trondheim Papers in Applied Linguistics Vol. IV :154-180

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

HOEY M. 1991b "Another perspective on coherence and cohesive harmony." in E. Ventola (ed.) 1991: 385-413

HOEY M. 1993 "A common signal in discourse: How the word *reason* is used in texts." in J. McH. Sinclair et al. (eds) 1993 :67-82

HOEY M. and WINTER E.O. 1986 "Clause relations and the writer's communicative task." in B. Couture (ed.) Functional Approaches to Writing: Research Perspectives London: Frances Pinter

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 psychological approach to academic writing.' in Research in English Language Teaching Vol. 3/1 :58-69

HUDSON R. 1984 Word Grammar Oxford: Basil Blackwell

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.

HUTCHINS J. 1977 "On the structure of scientific texts." in University of East Anglia Papers in Linguistics Vol. 11/23-33

HUTCHINS J. 1985 "Some problems and methods of text condensation." in University of East Anglia Papers in Linguistics Vol. 19/38-54

HUTCHINS J. 1987 "Summarization: some problems and methods." 151-173 in E. Jones (ed.) 1987 :Meaning: The Frontier of Informatics (Informatics 9) Proceedings: Kings College Cambridge 26-27 March 1987

HYMES D.H. 1971 On Communicative Competence Philadelphia: University of Pennsylvania Press

IDE. N.M. 1983 "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

JANNSEN S. "Automatic sense disambiguation with LDOCE: Enriching syntactically analysed corpora with semantic data." 105-135 in J. Aarts and W. Meijs (eds) 1990

JENNINGS E.M. 1990 "Paperless writing revisited." in Computers and the Humanities Vol.24 :43-48

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

JOHANSSON S. and A.B. STRENSTRÖM 1991 (eds.) English Computer Corpora Berlin: Mouton de Gruyter

JOHNS A.M. 1988 "Reading for summarizing: an approach to text orientation and processing" in Reading in a Foreign Language 4/2: 79-90

JOHNS A. M. and MAYES P. 1990 "An analysis of summary protocols of university ESP students." in Applied Linguistics Vol.11/3 :253-271

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

JONES S. and SINCLAIR J. McH. 1974 "English lexical collocations." in Cahiers de Lexicologie Vol. 24 :15-61

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

JUSTESON J.S. and KATZ S.M. 1991 "Co-occurences of antonymous adjectives and their contexts." in Computational Linguistics Vol.17/1 :1-19

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

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

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

KELLER M. and LLOYD P. 1994 (eds) Keywords in Evolutionary Biology New York: Harvard University Press

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

KHURSHID A. 1979"On abstracts and abstracting." in Annals of Library Science and Documentation Vol. 26 :14-20

KIERAS D.E. and BOVAIR S. 1981 "Strategies for abstracting main ideas from simple technical prose" Technical report No.10, University of Arizona.

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 KEENAN J. 1973 "Reading rate and retention as a function of the number of propositions in the base structure of sentences." in Cognitive Psychology Vol. 5 :257-274

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. 1987 "Aspects of English collocations." in W. Meijs (ed.) 1987 :133-140

KJELLMER G. 1990 "Patterns of collocability." in J.Aarts and W. Meijs (eds) 1990 :163-178

KJELLMER G. 1991 "A mint of phrases." in K. Aijmer and B. Altenberg (eds) 1991 :111-127

KJELLMER G. 1993 "Multiple meaning and interpretation: the case of *sanction*." in Zeitschrift für Anglistik und Amerikanistik Vol. 2/2 :115-123

KNORR-CETINA K. D. 1983 (ed.) 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

KOULOPOULOS T. M. 1992 "Document clustering." in Byte Magazine June :272-273

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. "The process of compilation." in J.McH. Sinclair (ed.) 1987 :62-85

KUCERA H. and FRANCIS W. N. 1967 Computational Analysis of Present Day American English Providence: Brown University Press

KUKULSKA-HULME A. (forthcoming) "Effective Knowledge Transfer: a Terminological Perspective." unpublished Ph.D thesis, Modern Languages Department: Aston University

KUZANWA N. B. 1987 "On the discourse structure of textbook review articles in language teaching journals: A case study of 'ESP' and 'ELT' journals." Unpublished MSc. thesis. Language Studies Unit, Aston University.

KYTÖ M, IHALAINEN O. and RISSANEN M. (eds.) 1988 Corpus Linguistics Hard and Soft Amsterdam: Rodopi
M.T. Trimble et al.(eds.) 1978 :53-73

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

LAURÉN C. and NORDMAN M. 1992 "Corpus selection in LSP research." in Schröder (ed.) 1992 :218-230

LAKOFF G. 1987 Women, fire and dangerous things. What categories reveal about the mind. University of Chicago Press: California

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

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

LEHR A. 1993 "Collocational analysis, from collocation theory of contextualism to computer-aided models." in Zeitschrift für Germanistische Linguistik Vol.21/1 :2-19

LEHRER A. 1974 Semantic Fields and Lexical Structure Amsterdam: North Holland Publishing Company

LEMKE J.L. "Text production and dynamic text semantics." in E. Ventola (ed.) 1991 :23-37

LEONOV V.B. and SERGEEVA N.E. 1980 "Evaluation of abstract clarity." in Scientific and Technical Information Processing Vol. 4 :33-42

LEVINSON S.C.1983 Pragmatics Cambridge: Cambridge University Press

LÉVI-STRAUSS C. 1962 La Pensée Sauvage Paris: Plon

LEVIN B. 1991 "Building a lexicon. The contribution of linguistics." in International Journal of Lexicography Vol. 4/3 :205-226

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

LJUNG M. 1991 "Swedish TEFL meets reality." in S. Johansson and B. Strenström (eds.) :245-256

LORCH F.R. and PUGZLES-LORCH E. 1986 "On-line processing of summary and importance signals in reading." in Discourse Processes Vol. 9/4 :489-497

LOVE A. 1993 'Lexico-semantic features of geology textbooks'. in English for Specific Purposes Journal Vol.12/3 :197-218

LOVERIDGE R. 1989 "Triangulation" unpublished Aston Business School methodology guide for post-graduates, Aston University.

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

LYNE A.A. 1975 "A word-frequency count of French business correspondance." in IRAL Vol. 13/2 :95-110

LYONS J. 1970 New Horizons in Linguistics Harmondsworth: Penguin Books

LYONS J., COATES R., DEUCHAR M. and GAZDAR G. (eds) 1987 New Horizons in Linguistics 2 Harmondsworth: Penguin Books

MACKIN R. (ed.) 1973 English Studies Series: Chemistry Oxford University Press

MAEDA T. 1981 "An approach to functional text structure analysis of scientific text and technical documents." in Information Processing and Management Vol. 17 :329-339

MAINGUENEAU D. 1987 Nouvelles Tendences en Analyse du Discours Paris: Hachette Université

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

MAKKAI A. 1992 "The challenge of the virtual dictionary and the future of linguistics." in International Journal of Lexicography Vol. 5/4 :252-269

MAKKAI A. 1988 "How to put the pieces of a poem together." in M. Ghadessy 1988 :145-159

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

MANN W. C. and THOMPSON S.A. 1986 "Relational propositions in discourse." in Discourse Processes Vol. 9/1 :57-90

MANN W. C. and THOMPSON S.A. 1988 "Rhetorical structure theory: Toward a functional theory of text organization." in Text Vol. 8/3 :243-281

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

MARTIN W. A.L B. and VAN STERKENBERG P. 1983 "On the processing of a text corpus." in R.R.K. Hartmann (ed.) 1983 Lexicography: Principles and Practice London: Academic Press :77-87

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

MAYES P.B. 1978 "A comparison of the readability of synopses and original articles for *Engineering Synopses*." in Journal of the American Society for Information Science Vol. 29 :312-313

McCAWLEY J.D. 1982 "How far can you trust a linguist?" in T.W. Simon and R.J. Scholes (eds.) Language, Mind and Brain. , London: Erlbaum :75-87

McCARTHY M. and CARTER R. 1994 Language as Discourse. Perspectives for language teaching New York: Longman

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

MEIJS W. 1987 (ed.) Corpus Linguistics and Beyond Amsterdam: Rodopi

MEIJS W. 1992 "Computers and Dictionaries" in C. Butler (ed.) :141-165

MEMET M. 1986 "L'abstract: outil de communication et technique d'apprentissage." in Langues Modernes Vol. 80/6 :53-57

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. 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. 1987 "The analysis of meaning." in J. McH. Sinclair (ed.) 1987 :86-103

MOORE J.D. and POLLACH M.E. 1994 "A problem for RST: The need for multi-level discourse analysis." in Computational Linguistics Vol. 18/4:537-544

MORRIS J. and HIRST G. 1991 "Lexical cohesion computed by thesaural relations as an indicator of the structure of a text." in Computational Linguistics Vol. 17/1 :21-48

MOSKOVICH 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 Lexicale Paris: Librairie Klincksieck

MULLER C. 1977 Principes et Méthodes de Statistique Lexicale Paris: Hachette Université

MURPHY G.L. 1990 "Noun phrase interpretation and conceptual combination." in Journal of Memory and Language Vol.29/1: 259-288

MYERS G. 1988 "The social construction of science and the teaching of English: An example of research." in English Language Teaching Documents 'Academic writing process and product' :143-150

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 Journal Vol. 11 :3-17

NAKAMURA 1991 "A study of the structure of the Brown corpus based upon the distribution of its vocabulary items." in Journal of Foreign Languages and Literature Vol.2.

NAKAMURA J. 1993 "Statistical methods and large corpora. A new tool for describing text types." in Baker et al. (eds.) 1993 :293-312

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

NEWMARK P. 1988 A Textbook of Translation Hemel Hempstead: Prentice Hall

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

OPPENHEIM R. 1988 "The mathematical analysis of style: a correlation-based approach." in Computers and the Humanities Vo.22 :241-253

OSBORNE G. 1992 "Computational analysis of idiomatic phrases in modern English." Unpublished M. Phil. Thesis, English Department, Birmingham Universitty

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

PALMER J. 1991 "Scientists and information: using cluster analysis to identify information style." in Journal of Documentation Vol. 47/2 :105-129

PALMER J. 1968 The Selected Papers of J.R. Firth 1952-59 Longman: London

PAPEGAAIJ and SCHUBERT R. 1988 A Corpus-based Bilingual Knowledge Bank for Distributed Language Translation DLT Publications Amsterdam.

PARSONS G. 1991 "Cohesion coherence: Scientific texts." in E. Ventola (ed.) 1991 :415-429

PATTEN T. 1992 "Computers and natural language processing." in C. Butler (ed.) :29-52

PAVEL S. 1993a "Neology and phraseology as terminology-in-the-making." in H.B. Sonneveld & 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. Canada

PAVEL S. and BOILEAU 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. 1985 "Two puzzles for linguistic theory: naturelike selection and naturelike fluency." in Richards and Schmidt (eds.) 1985 Language and Communication London: Longman

PETTINARI C. 1982 "The function of a grammatical alteration in 14 surgical reports." in W. Frawley (ed.) 1982: 145-183

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

PETERS A. 1989 The Limits of Language Acquisition Cambridge: Cambridge University Press

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 Exeter University Department of Linguistic and International Studies Monographs.

PICOCHE J. 1992 Précis de lexicologie française. L'étude et l'enseignement du vocabulaire. Paris: Nathan

POLLOCK J.J. and ZAMORA A. 1975 "Automatic abstracting research at Chemical Abstracting Service." in Journal of Chemical Information and Computer Sciences Vol.15/4 :226-232

POLSSKAYA O.B. 1986 "Improving the content structure of abstracts on the basis of queries from workers in new terminology." in Scientific and Technical Information Processing Vol. 12/5 :16-21

POTTER R. G. 1991 " Statistical analyses of literature: a retrospective on *CHum*: 1966-1990" in Computers and the Humanities Vol. 25 :401-429

PROCTER P. (ed.) 1992 The Cambridge Language Survey Prospectus Cambridge: Cambridge University Press

PROPP V. 1968*1928 The Morphology of the Folktale University of Texas Press

PUSTEJOVSKY J. 1991 "The generative lexicon." in Computational Linguistics 17/3: 46-53

QUIRK R. 1984 "Recent work on adverbial realisation and position." in J. Aarts and W. Meijs (eds.) 1984 :185-192

QUIRK R. 1992 "On corpus principles and design." in J. Svartvik (ed.) 1992 :457-469

QUIRK R., GREENBAUM S., LEECH G. and SVARTVIK J. 1985 A Comprehensive Grammar of the English Language London: Longman

RADA R., MILI H., LETOUREAU G. and JOHNSTON D. 1988 "Creating and evaluating entry-terms." in Journal of Documentation Vol. 44/1 :19-41

RADIEVSKAYA T.V. 1986 "Texts of abstracts considered in a linguopragmatic aspect." in Automatic Documentation and Mathemantical Linguistics Vol. 20/4 :53-63

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

REED A. and LAURIE SCHONFELDER J. 1979 "CLOC: a general-purpose concordance and collocations generator." in D.E. Ager , F.E. Knowles and J. Smith (eds.) :59-72

REEVES H. 1981 Patience dans l'Azur: L'Evolution cosmique Seuil: Paris

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. 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

RIGGS F.W. 1989 "Terminology and Lexicology: Their complementarity." International Journal of Lexicology Vol. 2/2 :89-109

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. 1977 "The notion of difficulty in science writing." Unpublished PhD. Thesis, Department of English, University of Birmingham

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

ROTH R.J. 1956 "How readable is chemical literature?" In American Documentation Vol.7 :215-221

RUMELHART D.E., McCLELLAND J.L. and the P.D.P Research Group 1986 "Parallel Distributed Processing." in Explorations in the Microstructure of Cognition Vol.1 77-109 Massachussets Institute of Technology

RUNDELL M. and STOCK P. 1992 "The corpus revolution." English Today April-October 1992

RUSH J.E., SALVADOR R. and ZAMORA A.V. 1971 "Automatic abstracting II: Production of indicative abstracts by application of contextual inference and syntactic coherence criteria." in Journal of the American Society for Information Science Vol. 22 :260-274

RUSSON D. (ed.) 1993 Current Serials Received Wetherby: British Library Document Supply Centre

RUSSON D. (ed.) 1992 Guide to Current British Periodicals Wetherby: British Library Document Supply Centre

SADLER V. 1989 Working with Analogical Semantics: Disambiguation Techniques in Distributed Language Translation Dortrecht: Foris Publications

SAGER J.C. 1990 A Practical Course in Terminology Processing Amsterdam: John Benjamins

SAGER J. C. 1991 "A theory of text production, modification, reception." in H. Schröder 1991 (ed.) : 34-57

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

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

SALAGER-MEYER F. 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

SAVOCA G. 1990 "A literary lexicography project for the Italian language." in Computers and the Humanities Vol. 24 :367-373

SCIENCE CITATION INDEX 1993a Permuterm Subject Index Institute for scientific information: Philadelphia

SCIENCE CITATION INDEX 1993b Journal Citation Reports Institute for scientific information: Philadelphia

SCHÄFFNER C., SHREVE G.M. and WIESEMANN U. 1987 "A procedural analysis of argumentative political texts." in Zeitschrift für Anglistik und Amerikanistik Vol. 35/2 :105-117

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

SCHRÖDER H. (ed.) 1992 Subject-oriented Texts: Languages for Special Purposes and Text Theory Berlin: Mouton de Gruyter

SCHUBERT K. 1986 Distributed Language Translation Amsterdam: Elsevier Science

SCOTT W.A.H. 1991 Chemistry Glasgow: Harper Collins

SCOTT D. and BRIDLE J. (rappporteurs) 1993 "Information handling" in Ostler N. (ed.) Paradigm Shift in Speech and Language Technology: Integrating with other media Proceedings of the SALT workshop, Camden Town 7-8 January 1993

SCOTT W.A.H. 1991 Chemistry Glasgow: Harper Collins

SEARLE J.P. 1969 Speech Acts London: Oxford University Press

SELINKER L., TARONE R. and HANZELI V. (eds.) 1981 English for Academic and Technical Purposes: Studies in Honor of Louis Trimble Newbury House: Mass. USA

SHARP B. 1989 "Elaboration and Testing of New Methodologies for Abstracting" unpublished Ph.D thesis, Modern Languages department, Aston University

SHERRARD 1986 "Summary writing: a topographical study." in Written Communication Vol. 3/3 :324-343

SHERRARD 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 Birmigham , 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.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. 1993a "Retrieving collocations from text: Xtract." in Computational Linguistics Vol19/1 :143-177

SMADJA F. 1993b "Xtract: an overview." in Computers and the Humanities No. 26: 399-413

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

SPERBER D. and WILSON D. 1986 Relevance: Communication and Cognition Oxford: Blackwell

STEMBERGER N. 1985 The Lexicon in a Model of Speech Production New York: Garland Press

STOTESBURY, H. 1990 "Finnish History Students as 'Liminal' Summarizers on the Threshold of Academia." Ph.D thesis, Kopi-Jyvä Oy: Reports from the Language Centre for Finnish Universities

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

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. and NAJJAR H. 1987 "The writing of research article introductions." in Written Communication Vol. 4:175-192

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

TARASOVA T. 1993 "Non-verbal elements in scientific text." Unpublished PhD 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. 'Oncogenes and cancer.' in J. Waxman and K. Sikera 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

TRIMBLE M.T., TRIMBLE L. and DROBNIC K. 1981 (eds.) English for Specific Purposes: Science and Technology Oregon State University: Corvallis

TYMA D 1981 "Anaphoric functions of some demonstrative noun phrases in EST" in Selinker et al. (eds.) :65-75

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 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.

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

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

WAITHE P. (ed.) Standard periodical directory. 16th edition Oxbridge Communications Inc.: Baltimore

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

WIDDOWSON H.G. 1977 "Description du langage scientifique." in Le Français dans le Monde No. 129 :15-21

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

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

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

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.O. 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.O. 1978 "A look at the role of certain words in information structure." in K.P. Jones and V. Horsnell (eds.) Informatics 3 London: ASLIB

WINOGRAD T. and FORES I. 1986 Understanding and cognition: A new foundation for design. Ablex Publishing: New Jersey.

WITTGENSTEIN L. 1957 Philosophical Investigations Oxford: Blackwell

WODAK R. 1990 "Discourse analysis: Problems, findings, perspectives." in Text Vol. 12/1-2: 125-132

WOOD P. 1982 "An examination of the rhetorical structures of authentic chemistry texts." in Applied Linguistics Vol. 3 :121-143

WÜSTER E. 1968 Enciklopedia Vortaro: Internationale Sprachnormung das Verhältniswort in Esperanto UEA, Rotterdam

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

YOUNG D.J. 1980 The Structure of English Clauses London: Hutchinson

YUMIN C. 1986 "An attempt at analysing English style." in J.Aarts and W. Meijs 1986 :219-227

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

ZIPF G.K. 1932 Studies of the Principle of Relative Frequency in Language Harvard University Press

APPENDIX A

THE PHARMACEUTICAL SCIENCES CORPUS (PSC) REFERENCE LISTS.

Journals are alphabetically listed according to the Science Citation Index (SCI) mnemonic code (CCP, CL etc) and not according to title. The Journal's rank in the SCI (1988) impact factor table (compared with 1 000 other journals) is listed as an approximate indicator of prestige. The relative size of the journal as a percentage of the PSC corpus is also noted. The Unix word count has been used for this, where the total corpus is of 150 papers, and 519 201 running words.

Papers are listed accoding to code (CL3 etc). For each paper one of several field classifications is noted (generally: cancer research / medicinal chemistry / pharmacology / structural chemistry). Only asterisked authors 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 an endogenous 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]pyreneinduced 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-Induced 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-Induced 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
Author: FC, etc al. Source: Int J. Cancer 53/4 1992

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 Medicinal 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]

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 ß-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: IH NMR 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

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

APPENDIX B

Below is a summary of the prospection and encapsulation codes as they have been described in Chapter 10 in the thesis. Codes are marked next to each sentence in each of the ten sample texts below. In order to make the analysis clearer, lexical rephrasings (coded E-lr) have been marked in **bold**, explicit discourse signals (coded E-x) have been underlined.

Note that in clause complexes, the first clause is referred to as a) and the second as b) and that if subordinate clauses have different postures this is marked in the Appendix. As set out in Chapter 10, internal encapsulation in clause complex is further subclassified; often as explicit signalling (coded E-x). The embedded or dependent clause has the same posture as the main clause where this is not differentiated in the Appendix.

TI -Text initial [No posture]

E - Encapsulation
[The default is 'logical implicit' and is not marked. Internal encapsulation is further subclassified]

- **x** - Explicit
- **e** - Ellipsis
- **d** - Deictic, including one of the following:
  - **ls** - Lexical refocussing [also known as 'selective']
  - **lr** - Lexical rephrasing
  - **i** - Including

P - Prospection
- **ts** - Topic selection.
- **at** - Attribution.
- **al** -Advance labelling.

P2 - The sentence is previously prospected

VE - Verbal Echo

O - Overlay

APPENDIX B1

Posture in structural chemistry article CC (author *SF*).

TITLE
TI
$1 Bioreversible Protection for the Phospho Group: Chemical Stability and Bioactivation of Di(4acetoxybenzyl) Methylphosphonate with Carboxyesterase.

ABSTRACT
E- ve ($1)
$2 In contrast to high chemical stability (T1/2 55.4 h at 36.4 oC), with porcine liver carboxyesterase the title compound (1) spontaneously decomposes first to the monoester(2) then to methylphosphonate, both reactions proceeding via the 4-hydroxybenzyl intermediates (3) and (4).

INTRODUCTION
a) TI
b) Ex
$3 Several compounds containing the phospho group [-PO(OH)2] are of therapeutic interest however at physiological pH, they are charged and many have not achieved their therapeutic potential principally because of poor transport across cell membranes.1

E-ls
$4 In an attempt to improve delivery, in some studies, their negative charges have been masked with the preparation of prodrug phosphoesters [-PO(OR)2], whose increased lipophilicity should facilitate transport into cells by passive diffusion.

a) E-ls
b) Ex
$5 The prodrug is required to liberate parent drug and for simple R groups, the first ester could possibly be removed by hydrolysis however the second ester is usually very resistant to cleavage.2

a) E-lr
b) P-at ('towards this end' = **E-lr**)
$6 For **bioactivation** to be successful, the R group must be metabolically labile and towards this **end** the di(acyloxymethyl) esters of the phospho group [-PO(O-CH2-OC(O)R')2] have been reported.3,4

a) P2
b) E-
$7 Srivastva and Farquar have investigated the di(acyloxymethyl) derivatives of benzyl and phenyl phosphate,3 which in the presence of porcine liver carboxyesterase readily decompose to the mono(acyloxymethyl) esters.

E-i
$8 This is thought to proceed by loss of the acyl group to give the hydroxymethyl derivative, followed byspontaneous loss of formaldehyde.3

a) E-x
b) E
$9 Although some benzyl or phenyl phosphate is released, the second acyloxymethyl group is removed only slowly by the esterase.

a) E
b) E-ls
$10 Compounds bearing a charge are reported to be poor substrates for esterases,5 and the slower rate for the removal of the **second acyloxymethyl group** is likely to be attributable to the inability of carboxyesterase to tolerate the anionic charge in close proximity to the active site.

a) E - ve ($10)
b) E
$11 The low reactivity of the monoanion, coupled both with the chemical instability of the acyloxymethyl esters3 and the potential problems with formaldehyde release have led us to consider alternative metabolically labile protection suitable for the phospho group.

E- (Possible P?)
$12 Benzyl esters of carbamates are useful prodrugs for compounds containing the amino group. 6,7

P2 or VE ($10, 13, 14,)
$13 In this report the use of carboxyesterase-susceptible dibenzyl phosphodiesters as prodrugs for the phospho group is explored.

a) P-att
b) P- ts
c) P2
$14 To promote the removal of the second ester, it is likely that the site of esterase attack needs to be well-separated from the monoanionic phospho group.

E-lr
$15 This **rationale** led us to explore the 4-acetoxybenzyl derivatives, in which for the monoanions, the charge is now nine bonds removed from the site of esterase attack, an increase of ~2.7 over the acyloxymethyl analogue.

a) E
b) E-lr and x ('so that')
$16 Methylphosphonic acid [MePO(OH)2] was chosen as a model compound, so that these **ideas** can be readily applied to the antiviral phosphonoacetate,8 and pamidronate (APD) used in the treatment of bone metastases9, both of which have poor oral bioavailability.

METHODS

E-ve ($15)
$17 Di(4-acetoxybenzyl) methylphosphonate (1) was prepared in 25 % yield from the reaction of two equivalents of 4- acetoxybenzyl alcohol with methylphosphonic dichloride in the presence of triethylamine.

E-
$18 The low-melting solid was purified by flash column chromatography and was fully characterised. In CDCl3, the 1H n.m.r. spectrum included dH 5.01 (2H, dd, Jgem 11.9 Hz, JPH 9.1) and 4.91 (2H, dd, Jgem 11.9, JPH 8.5) for the non-equivalent protons of the POCH2Ar groups.

E-ve ($11 ?)
$19 To test for chemical stability, a 5 mM solution of the diester (1) in potassium phosphate buffer (0.1 M, D2O, pD 8.0) - CD3CN (9:1, v/v) was monitored by 31P and 1H n.m.r. spectroscopy at 36.4 oC.

E-
$20 The diester gave dP 36.5 ppm and dH 7.31 (4H, d, JHH 8.5), 7.05 (4H, d, JHH 8.5), 4.92 (4H, d, JPH 9.2), 2.24 (6H, s) and 1.55 (3H, d, JPH 17.5).

E-x
$21 In contrast to the non-equivalence observed in CDCl3, the methylene protons now appear equivalent.

E- ve ($20)
$22 The hydrolysis of the diester was slow with a half life of 55.4 h ($k = 1.251 \pm 0.010 \times 10\text{-}2\ h\text{-}1$).

E-ls
$23 The formation of 4-acetoxybenzyl methylphosphonate (2) was confirmed by the synthesis of the sodium salt from the reaction of the diester (1) with sodium iodide,10 data on which included dP 27.5 ppm

and dH 7.41 (2H, d, JHH 8.6), 7.08 (2H, d, JHH 8.6), 4.81 (2H, d, JPH 7.2), 2.25 (3H, s) and 1.19 (3H, d, JPH 16.4).

E-lr
$24 As supported by hydrolysis of this **standard**, (2) decomposes further to methylphosphonate, dP 24.4 ppm (s, 1H decoupled) (q, JPH 16.4, 1H coupled), and dH 1.15 (3H, d, JPH 16.4) with a half life of 153.2 h (k = 4.525 ± 0.021 x 10-2 h-1).

E-lr
$25 **Other products** formed were acetate [dH 1.80 (3H, s)] and 4-hydroxybenzyl alcohol [dH 7.20 (2H, d, JHH 8.5), 6.81 (2H, d, JHH 8.5) and 4.45 (2H, s)].

a) E-ls
b) E-x (which suggests that)
$26 4-Acetoxybenzyl alcohol was not formed [authentic sample gives dH 7.36 (2H, d, 8.5), 7.06 (2H, d, 8.5), 4.55 (2H, s) and 2.25 (3H, s)] which suggests that the degradation of (1) and (2) must proceed with hydrolysis of the acetoxy group to give the 4-hydroxybenzyl intermediates (3) and (4) respectively.

RESULTS

a)P - al
b) P2
$27 Two studies support our view that the electron-donating 4-hydroxyl group will assist in the breaking of the benzyl-oxygen bond to give either monoester (2) or methylphosphonate together with the resonance-stabilised 4-hydroxybenzyl carbonium ion.

a) P2 (also prospected from $27)
b) E-x
$28 Firstly, dibenzyl methylphosphonate (5) is completely stable under similar conditions of hydrolysis over 48h,11 however at 100 oC the reaction proceeds via the benzyl carbonium ion with C-O cleavage.12

P2 (also prospected from $27)
b) E-x
$29 Secondly, after 323 minutes, 4-methoxybenzyl diphenyl phosphate undergoes 88% solvolysis in methanol, whereas under identical conditions the 3-methoxy or unsubstituted benzyl analogues are completely stable.13

E-
$30 The slow reactivity of (1) towards chemical hydrolysis is in marked contrast to the high reactivity reported for di(acetyloxymethyl) phenyl phosphate, which has a half-life of only 193 minutes at 37 OC and pH 7.4.3

E-x
$31 In contrast to the chemical stability, in the presence of 50 units of porcine liver carboxyesterase(SIGMA), a 5 mM solution (1 ml) of the diester (1) in phosphate buffer (0.1 M, D2O, pD 8.0) - CD3CN (9:1, v/v) at 36.4 oC rapidly decomposed in less than 3 min to give the monoester (2), which after 2 h gave only methylphosphonate.

E-lr
$32 A similar **reaction** with the monoester (2), monitored by 1H n.m.r. spectroscopy (Figure 1), shows that, in contrast to the chemical hydrolysis, the 4-hydroxybenzyl intermediate (4) was detected and gave dP 27.4 ppm and dH 7.24 (2H, d, JHH 8.7), 6.81 (2H, d, JHH 8.7), 4.71 (2H, d, JPH 7.1) and 1.17 (3H, d, JPH 16.3).

E-ls
$33 In the presence of 100 units of enzyme, after 15.5 min all of the monoester (2) had been metabolised to the 4-hydroxybenzyl intermediate (4), which was shown to decompose to methylphosphonate with a half-life of 17 min (k = 4.14 ± 0.45 x 10-2 min-1).

DISCUSSION

a) P-ve ($1, $11 and others)
b) P2
$34 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.

P2 (or E-lr?)
$35 The lower reactivity of the monoester with carboxyesterase when compared with the diester, could be exploited to provide a sustained release of parent drug.

E-ls
$36 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.

E-lr
$37 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

E-x / ve ($25)
$38 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.

a) E-x
b) E
$39 Instead of reacting with water, the carbonium ion may be trapped by another nucleophile, and possibilities include the enzyme, products or buffer.

E-ve (or overlay? - $31)
$40 The reaction profile for the decomposition of triester (1) with carboxyesterase is very similar to that of monoester (2) (Figure 1).

a) E

b) E-Ir
$41 For (1), two equivalents of the carbonium ion are generated, which does not lower catalytic efficiency, suggesting that this **intermediate** does not react with enzyme.

E-Ir
$42 In a **related reaction** 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.

E-Ir
$43 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.

P-ts
$44 To investigate the involvement of the buffer, the reaction of (1) with 5 units of carboxyesterase was repeated using 0.01 M phosphate buffer.

a) P2
b) E-Ir
$45 At all time points >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.

a) P (-x)
b) E-
$46 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 1h.

E-
$47 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

E-ve (or O $30)
$48 The higher reactivity of 4-hydroxybenzyl phosphate suggests a change in mechanism, with the electron-donating hydroxy group promoting C-O cleavage.

E-
$49 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.

a) E-ve ($47)
b) E-Ir
$50 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.

APPENDIX B2

Posture in structural chemistry article JCPT9 (authors *SF, WI, AG, DN*).

TITLE

TI $1 Bioreversible Protection for the Phospho Group: Bioactivation of the Di(4-acyloxybenzyl) and Mono(4-acyloxybenzyl) Phosphoesters of Methylphosphonate and Phosphonoacetate[1].

ABSTRACT

TI $2 The di(4-acetoxybenzyl) ester of methylphosphonate (4, X=H, R=Me) and the di(4-acyloxybenzyl) esters of methoxycarbonylmethylphosphonate (4, X=MeOOC, R=Me, Et, n-Pr, iso-Pr, n-Bu or t-Bu) were prepared from the appropriate benzyl alcohol and phosphonic dichloride.

E-lr $3 At pD 8.0 and 37 ^{o}C, both **series of compounds** hydrolyse with half-lives greater than 24h to the corresponding mono(4-acyloxybenzyl) esters 5, X=H or MeOOC, R=Me, Et, n-Pr, iso-Pr, n-Bu or t-Bu which were prepared by treatment of the di(4-acyloxybenzyl) esters 4 with sodium or lithium iodide.

E-ls $4 The mono(4-acyloxybenzyl) esters 5, X=H, R=Me and 5, X=MeOOC, R=Me, Et, n-Pr, iso-Pr, or t-Bu undergo chemical hydrolysis to methylphosphonate (6, X=H) and methoxycarbonylmethylphosphonate (6, X=MeOOC) respectively, together with 4-hydroxybenzyl alcohol and the appropriate acylate anion.

E-ls $5 The rate of hydrolysis of the mono(4-acyloxybenzyl) esters decreases as the length and steric bulk of the acyl group increases, with half-lives ranging from ~150 h for the acetyl analogues to 2240 h for the pivaloyl derivative.

E-ls 6 The hydrolyses of the di- and mono(4-acyloxybenzyl) esters were catalysed by porcine liver carboxyesterase (PLCE), and in all cases the acylate anion was formed.

E-ls $7 The rate of enzymatic hydrolysis was most rapid for the 4-n-butanoyloxybenzyl and 4-iso-butanoyloxybenzyl analogues.

E-ls $8 The carboxymethyl ester of the phosphonoacetate analogues was not cleaved by PLCE.

E- VE ($2) - $9 The methylphosphonate generated from the reaction of 4, X=H, R=Me in the presence of esterase and $H_2{}^{18}O$, did not contain ^{18}O attached directly to phosphorus.

E-lr $10 These **results** suggest that both the chemical and enzymatic hydrolyses of the mono(4-acyloxybenzyl) esters and the PLCE-catalysed hydrolyses of the di(4-acyloxybenzyl) esters proceed *via* hydrolysis of the acyl group to give the acylate anion and the unstable 4-hydroxybenzyl esters.

E- $11 The electron-donating 4-hydroxy group facilitates the cleavage of the benzyl-oxygen bond with the formation of the 4-hydroxybenzyl carbonium ion (9), which readily reacts either with water or the phosphate buffer.

E- $12 The 4-acyloxybenzyl phosphoesters provide the first example of a protecting group which will enable the bioactivation of phosphonate prodrugs at rates appropriate to biological systems.

INTRODUCTION

TI $13 Drugs that are charged at physiological pH often have limited cellular penetration which necessitates large intravenous doses to achieve a therapeutic effect.[2]

E- $14 The antiviral agent phosphonoacetic acid is partially triionic at physiological pH with pKa's of 2.30 (P-OH), 5.40 (COOH) and 8.60 (P-OH),[3] and when administered orally (20 mg $kg^{-1)}$ to both rabbit and monkey, only 2% and 8% of the dose was absorbed, respectively.[4]

E-ve (\$13/14) \$15 The delivery of therapeutic agents is often improved by the design of prodrugs, which undergo a chemical or enzymatic transformation, within the target organ, to release the therapeutic agent.[2]

a) E-ls
b) E-x
\$16 One approach to prodrug design involves the conversion of an active hydrophilic drug into an inactive lipophilic molecule, thus facilitating passive diffusion through cell membranes and other physiological barriers, for example the blood-brain barrier.

E-ls \$17 For drugs containing the phospho group (RPO_3^{2-}), neutral lipophilic phosphoesters [$RP(O)OR'_2$] are prodrug candidates.

a) P-al
b) P2
c) E-x
\$18 With simple R' alkyl analogues, the first **R' group** may be removed readily by chemical hydrolysis under physiological conditions however the second is usually very resistant to cleavage because the phosphorus of the anionic intermediate [$RP(O)(OR')O^-$] is unreactive towards nucleophilic attack.[6]

E- \$19 There is usually at least a million-fold decrease in the rate of removal of the second alkyl group, when compared with the first.

a) P- att b) P2 \$20 To facilitate release of the phospho group, one approach would be to design R' groups that do not require nucleophilic attack at phosphorus and subsequent P-O bond cleavage, for their removal.

E- \$21 Farquhar and co-workers[7,8] have examined a series of lipophilic di(acyloxymethyl) esters of benzyl and phenyl phosphate (1, R = Bn, Ph) as tripartite prodrug systems for the delivery of phosphates.

E-i \$22 These undergo bioactivation with esterase to give first the hydroxymethyl compounds 2, R = Bn, Ph which readily eliminate formaldehyde to give the diesters 3, R = Bn, Ph.

E- \$23 The rate of enzymatic hydrolysis can be controlled with different acyl groups, the more sterically hindered derivatives (R' = Bu^t) undergoing only slow hydrolysis.

a) E-x b) E-lr c) E- x \$24 Although the diesters 3 do degrade further to give benzyl and phenyl phosphate, this **second bioactivation step** is considerably slower than the first, presumably because the charged diester has a lower affinity for the esterase than the triester.

E- \$25 The acyloxymethyl esters of phosphonoformate have also recently been prepared.9

E-ls \$26 An approach to improve the monoanionic phospho intermediate as a substrate for esterases could be to distance the site of esterase attack from the anionic phospho group, by increasing the length of the linker.

E-x \$27 Instead of the acyloxymethyl group, we chose to explore the acyloxybenzyl group, which has been used for the delivery of amines as their carbamates.[10]

P- ts \$28 In this **study** we describe the metabolic activation of the model compound di(4-acetoxybenzyl) methylphosphonate (4, X=H, R=Me).

P2 \$29 Here, the charge on the monoanionic phosphonate intermediate 5, X=H, R=Me is nine bonds removed from the site of esterase attack, an increase of some 4 Å, the length of an aromatic ring plus one single C-C bond, over the acyloxymethyl analogues.

a) P-att b) P2 \$30 Previously we have shown that dibenzyl methoxycarbonylphosphonate, a triester of phosphonoformate was highly reactive towards chemical hydrolysis resulting in both P-O and C-P cleavage to give the diester and phosphite respectively.[11]

E-lr \$31 These **results** were in agreement with those of Krol et al.[12]

E-lr \$32 This **instability** presumably arises from the electron-withdrawing properties of the methoxycarbonyl group, suggesting that triesters of phosphonoformate might never be suitable prodrug forms.

E- lr \$33 In light of these **data**, the di(4-acyloxybenzyl) esters of the antiviral drug phosphonoacetate (4, X=MeOOC), in which the phosphorus atom and the carboxyl function are separated by a methylene group, were evaluated.

RESULTS AND DISCUSSION

P-ts \$34 The di(4-acetoxybenzyl) diester 4, X=H, R=Me and di(4-acyloxybenzyl) triesters (4, X=MeOOC, R=Me, Et, n-Pr, iso-Pr, n-Bu or t-Bu) were prepared by reaction of the appropriate benzyl alcohol[13] with methylphosphonic dichloride or methoxycarbonylmethylphosphonic dichloride[14] using a method similar to that previously described for the preparation of dibenzyl methoxycarbonylphosphonate.[11]

P2 \$35 The compounds were purified by flash column chromatography[15] and were fully characterised by elemental analysis and/or high resolution FAB mass spectrometry, infra-red spectroscopy, and ^{1}H, ^{31}P and ^{13}C NMR spectroscopy.

E-ls \$36 The ^{1}H NMR spectra in $CDCl_3$ showed that the benzylic protons were non-equivalent with the presence of 2 sets of doublets of doublets (J_{gem} ~12 Hz, J_{PH} ~9 Hz) at ~5 ppm.

E-x \$ 37 Interestingly, when the spectra were recorded using D_2O as solvent, the benzylic protons were equivalent giving rise to a doublet (J_{PH} ~9 Hz).

E- \$38 The salts of 4-acetoxybenzyl methylphosphonate (5, X=H, R=Me) and 4-acyloxybenzyl methoxycarbonylphosphonates (5, X=MeOOC, R=Me, Et, n-Pr, iso-Pr, n-Bu or t-Bu) were prepared in yields ranging from 35-81% by the action of sodium or lithium iodide on the appropriate di(4-acyloxybenzyl) ester 4, X=H or MeOOC adapting a published procedure.[16]

E- \$39 The diesters were characterised by high resolution FAB mass spectrometry, infra-red spectroscopy, and ^{1}H, ^{13}C and ^{31}P NMR spectroscopy.

E-ve (\$37) \$40 The ^{1}H NMR (D_2O) spectra all possessed a doublet at ~5 ppm (J_{PH} ~8 Hz) for the equivalent benzylic protons.

a) P - al b) P \$41 To evaluate the stability towards chemical hydrolysis, a solution of the diester 4, X=H, R=Me (5 mmol dm^{-3}) in potassium phosphate buffer (0.1 mol dm^{-3}, D_2O, pD 8.0) - CD_3CN (9:1 v/v) was monitored by ^{31}P and ^{1}H NMR spectroscopy at 36.4 ^{o}C.

E-ls \$42 The hydrolysis to potassium 4-acetoxybenzyl methylphosphonate (5, X=H, R=Me) was slow with a half life of 55.4 h ($k = 1.25 \pm 0.01 \times 10^{-2}\ h^{-1}$).

E- \$43 The monoester decomposed further to dipotassium methylphosphonate (6, X=H) with a half life of 153.2 h ($k = 4.52 \pm 0.02 \times 10^{-2}\ h^{-1}$).

a) E- b) E-x \$44 The small (3.6-fold) decrease in the rate constant for the hydrolysis of the monoester

when compared with the diester suggests that these compounds do not react by nucleophilic attack at phosphorus.

E-$45 The 4-acetoxy group should be susceptible to metabolic conversion by esterases to the electron-donating hydroxy group.

E-x (overides E-lr) $46 In contrast to this **moderate chemical stability**, the addition of PLCE (50 units) to a solution of the diester 4, X=H, R=Me (5 mmol dm^{-3}) in phosphate buffer (0.1 mol dm^{-3}, D_2O, pD 8.0) - CD_3CN (9:1 v/v, 1 ml) at 36.4 ^oC resulted in the rapid decomposition of 4 to give the monoester 5 in less than 3 minutes, which after 2 h gave only methylphosphonate (6, X=H).

a) E-lr b) E- $47 A similar **reaction** monitored by ^1H NMR spectroscopy showed that, in contrast to the chemical hydrolysis, in the presence of PLCE (100 units), the monoester 5, X=H, R=Me was completely metabolised within 15.5 min to the 4-hydroxybenzyl intermediate 8, X=H, which then decomposed to methylphosphonate (6, X=H) with a half-life of 17 min ($k = 4.14 \pm 0.45 \times 10^{-2}$ min^{-1}).

E-lr $48 The **formation** of potassium acetate and 4-hydroxybenzyl alcohol, and the absence of 4-acetoxybenzyl alcohol in the chemical and PLCE-catalysed reactions of 4 and 5, X=H, R=Me suggests that their degradation must first proceed with hydrolysis of the acetyl group to give the 4-hydroxybenzyl intermediates 7 and 8 respectively.

E-x $49 It is then proposed that the electron-donating 4-hydroxy group promotes cleavage of the benzyl-oxygen bond to give either monoester 5 or methylphosphonate (6, X=H) together with the resonance-stabilised 4-hydroxybenzyl carbonium ion (9).

E-lr $50 This **proposed mechanism of hydrolysis** is supported by the following evidence: (i) when the hydrolysis of diester 4, X=H, R=Me with PLCE (50 units) was repeated in the presence of 80% ^{18}O-enriched water, only singlets were observed by ^{31}P NMR spectroscopy for the intermediate 8, X=H and methylphosphonate (6, X=H).

a) E-lr b) P al b) P2... $51 This **result** confirms that i) the ^{18}O label is not attached to phosphorus in either compound, (ii) dibenzyl methylphosphonate is completely stable under similar conditions of hydrolysis over 48h,[11] however, at 100 ^oC, the reaction proceeds *via* the benzyl carbonium ion with C-O cleavage,[17] (iii) after 3.3 h, 4-methoxybenzyl diphenyl phosphate undergoes 88% solvolysis in methanol, whereas under identical conditions the 3-methoxy or unsubstituted benzyl analogues are completely stable,[18] and (iv) the attempted synthesis of di(4-methoxybenzyl) methoxycarbonylmethylphosphonate from 4-methoxybenzyl alcohol and methoxycarbonylmethylphosphonic dichloride gave only di(4-methoxybenzyl) ether.

E-lr $52 A similar **observation** was made during the attempted synthesis of the corresponding triester of phosphonoformate which suggests that P-benzyl esters substituted with *para* electron-donating substituents are unstable and decompose to give a benzyl carbonium ion.

a) P-at b) P $53 Although the products derived from the phospho group of the diester 4, X=H, R=Me are known, the fate of the benzyl group is more complex with only ~ 30% of the product derived from the proposed 4-hydroxybenzylcarbonium ion (9) being present as 4-hydroxybenzyl alcohol at early time points.

E-ls $54 It is possible that the carbonium ion may be trapped by nucleophiles other than water, and possibilities include the enzyme, products or buffer.

a) E- b) E-x $55 The catalytic efficiency of the esterase was not impaired during the hydrolysis, suggesting that this intermediate does not react with enzyme.

E-lr $56 In a related **reaction** 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.

E-ve ($53) $57 An analogous reaction of the 4-hydroxybenzyl carbonium ion with 4-hydroxybenzyl alcohol would give a trisubstituted ring, however, the ^1H NMR evidence only supports 1,4-disubstituted products.

E-ve ($53) $58 To investigate the involvement of the buffer, the reaction of 4, X=H, R=Me with PLCE (5 units) was performed using 0.01 mol dm^{-3} phosphate buffer.

a) E- b) P-att c) P2 $59 At all time points, >90% of the carbonium ion was trapped as 4-hydroxybenzyl alcohol which suggests that with the original 0.1 mol dm^{-3} buffer, inorganic phosphate can compete with water to trap the carbonium ion.

E- $60 Peaks in the NMR spectra of the reaction mixture with 0.1 mol dm^{-3} buffer at dp 3.72 ppm and d_H 4.64 (2H, d, J_{PH} 5.4), 6.81 (2H, d, J_{HH} 8.4) and 7.26 (2H, d, J_{HH} 8.4) ppm are for 4-hydroxybenzyl phosphate (10), which has an approximate half-life of 1 h.

E- $61 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.[21,22]

E-ls $62 The higher reactivity of 4-hydroxybenzyl phosphate suggests a change in mechanism, with the electron-donating hydroxy group promoting C-O cleavage.

E-ls $63 The ready removal of both 4-acetoxybenzyl groups from 4, X=H, R=Me with PLCE confirms that diesters of this type are bioreversible derivatives of the phospho group.

E-lr $64 This **approach** was then applied to the antiviral agent phosphonoacetate.

E- $65 The substrate specificity of esterases is dependent on the length and steric properties of the groups on either side of the ester link.[23]

E- ls $66 The potential to control the rate of esterase-catalysed hydrolysis with the nature of the acyl group was explored with the phosphonoacetate analogues 4 and 5, X=MeOOC, R=Me, Et, n-Pr, n-Bu, iso-Pr and t-Bu.

E-ls (P al?) $67 The chemical hydrolyses of the 4-acyloxybenzyl analogues of phosphonoacetate were examined first.

E- $68 A solution of the triesters 4, X=MeOOC (1 mmol dm^{-3}) in potassium phosphate buffer (0.1 mol dm^{-3}, D_2O, pD 8.0)-CD_3CN (9:1, v/v) at 37 ^oC were monitored by ^1H NMR spectroscopy.

E-x $69 Interestingly the rates were not very sensitive to the nature of the acyl group and further experiments are required to completely elucidate their hydrolytic profile.

E-ve ($66) $70 The hydrolyses of the diesters 5, X=MeOOC (5 mmol dm^{-3}) in phosphate buffer (0.1 mol dm^{-3}, D_2O, pD 8.0)-CD_3CN (9:1 v/v) at 37^oC followed typical first order reaction kinetics.

E-ls $71 By ^1H NMR spectroscopy, the diesters were shown to degrade to methoxycarbonylmethylphosphonate (6, X=MeOOC), 4-hydroxybenzyl alcohol and the acylate anion.

(P- al- Table) $72 The rate constants and half-lives are given in Table 1.

E-ls $73 The half-lives for the acetyl analogues of 5, X=H and MeOOC are both approximately 150 h, suggesting reaction by similar mechanisms.

E-ve ($60 etc.) $74 The rate of hydrolysis decreases as the length and b-alkylation of the acyl group

increases, the reactivity profile being comparable to that reported for the hydrolyses of ethyl acylates.[24]

E-ls $75 For all diesters 5, X=MeOOC, 4-hydroxybenzyl alcohol was formed, whereas the 4-acyloxybenzyl alcohols were not.

E-i $76 This is consistent with hydrolysis at the acyl group to give the 4-hydroxybenzyl intermediate 8, X=MeOOC.

P - at $77 Attention was then turned to the enzymatic bioactivation of the triesters 4 and diesters 5, R=MeOOC.

P2 $79 Porcine liver carboxyesterase is known to be a mixture of seven enzymes,[25] therefore a detailed study to determine K_m and V_{max} for the substrates seemed inappropriate.

E- $80 Each half-life is related to K_m and V_{max} by Equation 2 which indicates that a plot of $t_{1/2}$ against S_o' should be linear.

E-lr $81 In the PLCE experiments, this **analysis** showed a generally constant value for $t_{1/2}$ throughout the reaction.

E-i $82 This arises when $K_m >> S_o'$ which approximates the model to a first order degradation when the degradation rate constant is given by $k = V_{max} / K_m$.

E-x $83 Esterase hydrolyses have <u>thus</u> been analysed according to the first order kinetic model but $t_{1/2}$ values are quoted, rather than rate constants, to reflect this approximatio

E- $84 The phosphonoacetate triesters 4, X=MeOOC (1 mmol dm^{-3}) were incubated with PLCE [0.5 units (R=Me), 0.05 units (all other triesters)] in potassium phosphate buffer (0.1 mol dm^{-3}, pH 7.4)-MeCN (9:1 v/v, 1 ml).

E-ls $85 The reactions were monitored by isocratic ion-pair reversed-phase HPLC, using acetonitrile-10 mM tetrabutylammonium hydroxide in water.

(P - table) $86 The half-lives, reflecting K_m/V_{max}, are given in Table 2.

E-x $87 The reaction of di(4-acetoxybenzyl) methoxycarbonylmethylphosphonate (4, X=MeOOC, R=Me) (5 mmol dm^{-3}) with PLCE (5 units) was also monitored by ^1H and ^{31}P NMR spectroscopy.

E- $88 The triester was initially observed to degrade to the 4-acetoxybenzyl diester 5, X=MeOOC, R=Me [d_H including 2.26 (s, CH_3CO), 3.58 (s, OCH_3), 4.86 (d, J_{PH} 7.3 Hz, CH_2OP), 7.08 (d, J_{HH} 8.4 Hz, Ar) and 7.40 ppm (d, J_{HH} 8.4 Hz, Ar), and d_P 15.34 ppm (s)], 4-hydroxybenzyl alcohol [d_H 4.45 (s, CH_2), 6.81 (d, J_{HH} 8.6 Hz, Ar) and 7.19 ppm (d, J_{HH} 8.4 Hz, Ar)] and potassium acetate [d_H 1.80 (s)].

E-x $89 <u>In contrast to</u> the chemical hydrolysis, in the presence of PLCE (5 units), the pivaloyl triester 4, X=MeOOC, R=t-Bu (5 mmol dm^{-3}) degraded concomitantly to the diester 5, X=MeOOC, R=t-Bu [d_H including 1.27 (s) $C(CH_3)_3$] and potassium pivaloate [d_H 1.00 (s)].

E-lr $90 These **results** suggest that the PLCE-catalysed hydrolyses proceed *via* the 4-hydroxybenzyl intermediates 7, X=MeOOC, which then spontaneously degrade to the diesters 5, X=MeOOC.

E- $91 The phosphonoacetate diesters 5, X=MeOOC (5 mmol dm^{-3}) were incubated with PLCE (10 units) in potassium phosphate buffer (0.1 mol dm^{-3}, D_2O, pD 8.0)-CD_3CN (9:1 v/v, 1 ml).

E- $92 By ^{1}H and ^{31}P NMR spectroscopy the products were identified as dipotassium methoxycarbonylmethylphosphonate (6, X=MeOOC), data on which included d_H 2.59 (d, J_{PH} 19.6 Hz, PCH_2) and 3.58 ppm (s, OCH_3), and d_P 11.70 ppm (s), (t, J_{PH} 18.5 Hz, ^{1}H coupled)], and 4-hydroxybenzyl alcohol.

a) E- x b) E-x $93 In contrast to the chemical hydrolysis reactions, up to 20% of the 4-hydroxybenzyl intermediate 8, R=MeOOC was detected, data on which included d_H 2.75 (d, J_{PH} 20.5 Hz, PCH_2), 3.58 (s, OCH_3), 4.75 (d, J_{PH} 7.1 Hz, CH_2OP), 6.78 (d, J_{HH} ~8 Hz, Ar) and 7.23 ppm (d, J_{HH} ~8 Hz, Ar), and d_P 15.07 ppm (s), which confirms the formation of the monoester 6, X=MeOOC *via* the proposed 4-hydroxy promoted benzyl-oxygen bond cleavage.

a) E- b) E- x $94 Neither methanol [$d_H$ 3.25] nor phosphonoacetic acid [d_H 2.52 ppm (d, J_{PH} 20.1 Hz, PCH_2) and d_P 17.41 ppm (s)] were detected, showing that the monoester 6, X=MeOOC is not a substrate for the esterase, presumably because of its dianionic nature.

P (table) $95 The reaction profile for the PLCE-catalysed hydrolysis of 5, X=MeOOC, R=iso-Pr is shown in Figure 1.

E-lr $96 This **degradation** follows a 5->8->6 consecutive pathway with the concentration of the individual components (5_t, 8_t, 6_t) at time t being modelled by:

$$5_t = 5_0 \cdot \exp(-k_1 t)$$

$$8_t = \frac{5_0 \cdot k_1}{k_2 - k_1} \cdot [\exp(-k_1 t) - \exp(-k_2 t)]$$

$$6_t = 5_0 \cdot [1 - \frac{k_2 \cdot \exp(-k_1 t) - k_1 \cdot \exp(-k_2 t)}{k_2 - k_1}]$$

E-lr $97 Parameter estimation from these **equations** was undertaken by simultaneous non-linear regression of the time-concentration data collected during the degradation of 5.

E-ve ($90) $98 The 4-hydroxybenzyl intermediate 8 gave 6 with a half-life of 1.4 min.; some two-fold more reactive than the iso-propyl diester 5 which had a half-life of 3.1 min.

E-x $99 The intermediate 8, X=MeOOC was also significantly less stable than the corresponding methylphosphonate analogue 8, X=H ($t_{1/2}$ 17 min.), presumably by virtue of the electron-withdrawing methoxycarbonyl substituent favouring loss of the charged phosphonate.

E-ve ($88) $100 The half-lives for the hydrolyses of the 4-acyloxybenzyl diesters 5, X=MeOOC catalysed by PLCE are given in Table 2.

E-ls $101 For the straight chain acyl compounds, the acetyl analogues (R=Me) of the triesters (4, X=MeOOC) and diesters (5, X=MeOOC) are considerably less reactive than the longer chain derivatives (R=Et, Pr).

E-lr $102 This **result** is in agreement with those reported for the horse liver carboxyesterase-catalysed hydrolyses of ethyl acylates,[23] and the PLCE-catalysed hydrolyses of phenyl acylates.[28]

E-ls $103 The active site of PLCE appears to optimally accommodate a four carbon acyl group, with the greatest rate of hydrolysis being for the n-butanoyl and iso-butanoyloxybenzyl triesters 4 and diesters 5, X=MeOOC, R=n-Pr, iso-Pr (Table 2).

a) E-ls b) E-x $104 Branching of the acyl chain of the ester increases the affinity ($1/K_m$) but decreases

the reactivity (V_{max}) with horse liver esterase.[23]

E-i or E-ls $105 Here similar half-lives are observed for the 4-pivaloyloxybenzyl esters, 4 and 5, R=t-Bu and the corresponding straight chain analogues R=n-Bu, presumably resulting from a balance of these factors.

E-ve (103) $106 The esterase-catalysed reactions of the diesters 5, X=MeOOC utilised a 40-fold increase in the amount of PLCE when compared with the triesters 4, X=MeOOC.

E-o ($32) $107 The diesters are much poorer substrates than the triesters, which is likely to be attributable to the anionic nature of the diester.

E-x $108 In support, Levy and Ocken found that ethyl potassium succinate was neither a substrate nor inhibitor of PLCE, whereas diethyl succinate is a good substrate.[29,30].

a) E-i b) E-i $109 They attributed this to the proposal that charged compounds are unable to form a Michaelis-complex with the enzyme, and suggested that this could be due to charge repulsion because the catalytic site can only be reached through a cluster of negatively charged groups on the enzyme.

a) E-lr b) P -at $110 The potential application of this **bioreversible protecting group** for the delivery of phosphates and phosphonates to the brain prompted us to investigate the plasma stability and bioactivation by brain S9 fraction of some of the compounds.

P $111 The di(4-acyloxybenzyl) triesters 4, X=MeOOC, R=Me and n-Pr were incubated with human plasma under physiological conditions (pH 7.4, 37°C) and the reactions were monitored by HPLC.

E-ls $112 The triesters degraded to the diesters 5, X=MeOOC, R=Me and n-Pr with approximate half-lives of 6.5 and 8.5 min respectively.

E-ls $113 The mono(4-acyloxybenzyl) diesters 5, X=MeOOC, R=Me, n-Pr and t-Bu were incubated with human plasma and the reactions monitored by ^{31}P NMR spectroscopy.

E-ls $114 The acyloxybenzyl diesters 5 degrade in plasma to give methoxycarbonylmethylphosphonate (6, X=MeOOC).

E-ve (99) $115 The pivaloyl analogue (R=t-Bu) shows high stability with a half-life of 154 h, whereas the acetyl (R=Me) and butanoyl (R=n-Pr) analogues are considerably more reactive, with half-lives of 9.2 and 8.65 h respectively.

E-ls (refers to table) $116 Similar to the trends observed for chemical hydrolysis (Table 1), the rates of decomposition of the triesters 4 in plasma are faster than for the diesters 5.

E-ls $117 The rate of hydrolyses are similar for the 4-acetyl and 4-butanoyl derivatives, in contrast to the results with PLCE (Table 2) in which the 4-acetoxybenzyl diester 5, R=Me was much more stable than the 4-butanoyloxybenzyl derivative 5, R=n-Pr.

E-i $118 This suggests that human plasma carboxyesterases have different substrate specificities when compared to those in PLCE.

E- x $119 Again, the methoxycarbonyl group was unaffected by human plasma, confirmed by the stability of a sample of disodium methoxycarbonylmethylphosphonate (6, X=MeOOC) in plasma.

E- $120 Bioactivation of the 4-acyloxybenzyl diesters 5, X=MeOOC, R=Me, n-Pr and t=Bu with the S9 fraction of porcine brain was monitored by ^{31}P NMR spectroscopy.

E-ls $121 The diesters were found to degrade to methoxycarbonylmethylphosphonate (6, X=MeOOC) with half-lives of 2.0, 5.3 and 48 hours respectively.

E-ve (115) $122 The porcine brain-mediated bioactivation of the diesters follow a similar trend to that observed with human plasma, the pivaloyl analogue being considerably more stable than the acetyl or butanoyl analogues.

E-ls $123 The bioactivation of the triesters of phosphonoacetate to the methyl ester 6, X=MeOOC has been demonstrated with PLCE, human plasma and porcine brain S9 fraction.

E- ls $124 Further bioactivation to the parent drug has yet to be achieved.

P-at / P2 $125 Studies are underway to replace the methoxy group with a substituent that could be removed bioreversibly such as acyloxymethyl[2] or 4-acyloxybenzyl.

P or E-lr $126 A potential problem associated with **the 4-acyloxybenzyl prodrug approach** is the release of the highly reactive 4-hydroxybenzyl carbonium ion which may interact with cellular nucleophiles (eg. DNA, glutathione) and cause toxicity.[10]

a) E- x b) E-x $127 Indeed, when tested for antiviral properties, the 4-acyloxybenzyl- and 4-pivaloyloxybenzyl triesters, 4, X=MeOOC, R=Me and t-Bu, exhibited acute toxicity, possibly due to the generation of the carbonium ion.[30]

E-lr (1st reformulating and 2nd refocussing) $128 For this **reason** methods to trap this **intermediate** internally are being investigated.

E-lr $129 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,[7,8] strong acid[32] or trimethylsilylbromide.[33]

P -at (future studies?) $130 Studies are in progress to extend the 4-acyloxybenzyl prodrug system for the delivery of a range of phosphates, including the monophosphates of AZT and DDC.

APPENDIX B3

Posture in structural chemistry article JCPT8 (authors *WT, CS*).

(preformulations are marked in **bold**. Sentences are numbered by $)

TITLE

TI $1 Structural Studies on Bio-active Molecules. Part 17.Crystal Structure of 9-(2' Phosphonylmethoxyethyl)adenine (PMEA)1,§

ABSTRACT

TI $2 The crystal structure of the antiviral agent PMEA (1) has been determined to R = 0.032 and reveals a zwitterion whose side chain emerges in-plane from the adenine ring; the conformation is compared with that of crystal structures of acyclic nucleoside analogues and AMP (2).

INTRODUCTION

TI $3 The acyclic nucleotide analogue, 9-(2' phosphonylmethoxyethyl)adenine (1, PMEA) is currently undergoing further evaluation as a drug for the treatment of AIDS.2

E- i $4 It is a potent and selective inhibitor of the human immunodeficiency virus (HIV), with an ED50 of 2 mM in MT-4 cells.3

E- ls $5 PMEA has shown stronger in vivo antiretrovirus activity than AZT against Moloney murine sarcoma virus-induced tumour formation.4

E- ls $6 PMEA is also active against a broad range of herpes viruses, including cytomegalovirus.3,4

E- ls $7 Following penetration into the cells, bisphosphorylation of PMEA is catalysed by host kinases.5

E- ls $8 The bisphosphate of PMEA has been shown to inhibit DNA polymerase,6 ribonucleotide reductase,7 and interestingly, reverse transcripase.5

E- ve (3?) $9 As part of our efforts to synthesise prodrug forms, a sample of PMEA was prepared as both a starting material and as a standard for HPLC analysis.

E- ls $10 Crystals suitable for X-ray diffraction were formed, and it is the crystal structure of PMEA that is discussed.

METHODS

E-ve ($9) $11 The bis(trimethylsilyl) ester of PMEA was synthesised by the method described by Holy and Rosenberg.8

E- lr $12 From this **compound**, these workers formed the disodium salt however treatment of the ester with water gave the free acid of PMEA.

E- $13 Crystallisation from water gave colourless laths, which were fully characterised spectroscopically and by elemental analysis.

E- $14 The 1H NMR spectrum showed that the protons in each methylene group are equivalent, suggesting a flexible conformation of the "sugar moiety" in solution.

E- $15 Reflections were collected to 2qmax = 54o, yielding 2433 independent reflections (Rint = 0.020) of which 2196 were considered observed with [Fobs > 3s(Fobs)].

E- lr $16 The **structure** was solved by direct methods,13 methylene H atoms were placed in calculated positions, and all other H atoms were located in a difference electron density map.

E- $17 Full-matrix least-squares refinement13 of all atomic coordinates together with anisotropic thermal parameters for non-H atoms and isotropic temperature factors for H atoms converged at R = 0.032, Rg = 0.052.

E- $18 No peak on a final difference electron density map exceeded 0.33 eÅ-3.

E- ve (16) $19 Lengths and angles of covalent and hydrogen bonds, and observed and calculated structure factors have been deposited at the Cambridge Crystallographic Data Centre: see Instructions for Authors, J. Chem. Soc., Perkin Trans. 1, 1990, Issue 1.

RESULTS AND DISCUSSION

TI $20 The unit cell of PMEA is shown in the Figure and the atomic coordinates are given in the Table.

E- ls $21 PMEA was found to crystallise as the zwitterion protonated at N(1).

E- lr $22 **The most unusual feature of this structure** is the "glycosidic" torsion angle.

E- ls $23 In 10 molecules10 of acyclonucleosides for which C(1') is secondary this torsion angle averages 96, yet in PMEA C(8)-N(9)-C(10)-C(11) is only 3.3.

E- ls $24 The close 1,4-contact that arises between C(8) and C(11) is relieved by expanding angle C(8)-N(9)-C(10) to a value 7.4(1)o greater than C(4)-N(9)-C(10).

E- lr $25 While syn-anti conversion has been recognised in acyclonucleosides11, the present **structure** provides a concrete representation of the likely intermediate.

E- ve (23) $26 The remaining twist angles along the chain to the phosphonate group are gauche, trans, trans about C(10)-C(11), C(11)-O(12), O(12)-C(13) respectively.

E- ls $27 Bond distances and angles in the adenine moiety resemble those in the standard model of protonated adenine12 within 3s except in the vicinity of N(6), where a hydrogen bond to an adjacent phosphonate oxygen and another to an inversion-related N(7) may perturb the geometry.

E- ls $28 A strong [2.610(1) Å] intermolecular hydrogen bond links N(1) and phosphonate O(15), and the screw axis creates columns of adenine rings while the phosphonates occupy separate domains (Figure).

E- ls $29 P-O bond lengths are correlated with the nature and number of other attachments to the O atoms: the longest bond is to O(17), which is protonated; the next longest, to O(15), which accepts two hydrogen bonds; and the shortest, to O(16), which accepts one.

APPENDIX B4 __

Posture in structural chemistry article JCPT10 (author *WF, CS, HW.*).

(rephrasings are marked in **bold**. Sentences are numbered by $)

TITLE
TI Latent Inhibitors. Part 7. Inhibition of Dihydro-orotate Dehydrogenase by Spirocyclopropanobarbiturates.

ABSTRACT
TI $1A series of 5-spirocyclopropanobarbiturates bearing alkyl and aryl substituents on the cyclopropane ring has been synthesized.

E- lr $2 Dihydro-orotate dehydrogenase from C/ostridium oroticum was shown to be inhibited by these **compounds**.

E- lr $3 A **related series** of 5-membered-ring compounds (hydantoins and pyrazoles) was prepared but all the compounds were found to be inactive.

a) E-lr b) E- x $4 In order to correlate these **observations** with previous results concerning 5-arylmethylhydantoins and 5-arylidenehydantoins as inhibitors, 5-arylidenebarbiturates were also assessed as inhibitors and found to be the most active of the compounds investigated.

E- lr $5 The **results** are interpreted in the context of molecular recognition by this enzyme and the possibility of using substrate surrogates as templates for constructing latent inhibitors of enzymes.

INTRODUCTION
TI $6 Previous studies in this laboratory have shown that dihydroorotate dehydrogenase (DHODase), which is a significant target for chemotherapy,2 can be inhibited irreversibly by 5-arylmethylhydantoins, provided that a 1-carboxy-2-phenylethyl substituent is present at N-3. 1 3

E- x $7 We have also shown that a cyclopropane ring can be introduced into substrates to afford latent inhibitors of a wide variety of enzymes.4,5

E- ls $8The goal of investigating such latent inhibitors is to establish design methodologies that will make it possible to obtain highly selective enzyme-activated inhibitors as lead compounds for chemotherapy.

a) E- lr b) E- x $9 **Selectivity** can be generated by enzyme activation but since many naturally occurring compounds are substrates or products for more than one enzyme, substrates modified to contain latent functionalities may not attain the desired selectivity.

a) E- b) E- ls $10 In view of our fortuitous discovery of compounds that have a totally different molecular framework from the substrate, namely the hydantoins that inhibit DHODase, 1.3 the questiona rose whether substrate surrogates that differ markedly in overall structure from the normal substrate but which retain key features for enzyme-catalysed activation might provide a general design concept for highly selective enzyme inhibitors.6

E- lr $11 So far, such **studies** have been restricted largely to peptidases for which some substrate surrogate irreversible inhibitors based upon lactones have been described. 7

P-at $12 In this paper we describe the synthesis and evaluation of a number of heterocyclic compounds designed as potential inhibitors of DHODase to explore the potential of the concept of substrate surrogate in this case.

P2- DESIGN AND SYNTHESIS OF INHIBITORS.

TI $13 A preliminary study has shown that 5-spirocyclopropanobarbituric acid was an inhibitor of DHODase 6 and in view of the hydrophobic binding effects discovered with the hydantoins as inhibitors, the synthesis of a series of alkyl- and aryl-substituted 5-spirocyclopropano analogues was undertaken (Scheme I a).

E- $14 Condensation of aryl or alkyl aldehydes with diethyl malonate led to the ~"B-unsaturated diesters (la-h) which were cyclopropanated with trimethylsulphoxonium ylide to afford compounds (2a-h).

E- $15 The l,l-cyclopropanedicarboxylate diesters were cyclised to the corresponding barbituric acids (3a-g) by reaction with urea in the presence of potassium t-butoxide in yields between l l and 77%.

a) E- ve (16) b) E-x $16 The synthesis of heterocycles when using urea as a nucleophile is not usually a successful reaction; the acceptable yields obtained in this case are probably due to the cyclopropane ring which promotes cyclisation by restricting conformational freedom.

a) E- x b) E- lr $17 The cyclisation, however, failed in the case of the 4-chlorophenyl-substituted diester (2h), and to obtain this and other **aryl-substituted compounds** an alternative route was used (Scheme l b).

E- lr $18 A further **polarisation** of one of these groups by DHODase at its active site would enhance the reactivity of the small ring towards nucleophilic addition.

E- lr $19 The same **argument** can reasonably be applied to analogous five-membered-ring compounds such as pyrazolones.

E- lr $20 These **compounds** (5a and b) were readily available by condensation of the diesters (2) with hydrazine or substituted hydrazines (Scheme 2a).

a) E- b) E-lr $21 The hydantoins studied previously were shown to have their key chemical reactivity at the 5-substituent and a **further approach** to probe the surrogate substrate concept was to prepare a series of 5-cyclopropanohydantoins (7a-c) (Scheme 2b).

E- ls $22 5-Spirocyclopropanohydantoin (6) was obtained in three steps from diethyl cyclopropane-l,l-dicarboxylate .8

E- $23 Alkylation on N-3 was successful with l-bromo- 2-phenylethane in the presence of potassium t-butoxide to give compound (7b), but took place only in low yield with ethyl bromoacetate to give compound (7a) and failed completely with methyl 2-bromo-3-phenylpropanoate.

RESULTS AND DISCUSSION

P - ts Inhibition of Dihydro-orotate Dehydrogenase.

P2 (E- ve 13) $24 The inhibition of DHODase was studied under the standard conditions used previously 1,3 and all of the alkyl- and aryl-substituted spirocyclopropanobarbiturates (3a-h and j) were found to be irreversible inhibitors and were all bound more strongly and were more reactive than the parent unsubstituted compound (Table).

E-lr $25 The **most reactive compound** (3d), which bore an isobutyl substituent, was 3 times more reactive than the parent compound.

E-lr $26 The **tightest binding** was obtained with the n-butyl substituted compound (3e) which had ca. 10 times the affinity of the parent compound for the enzyme.

E- x $27 However, the range of affinity and reactivity is not large and clearly, for the range of substituents chosen, the enzyme is not very discriminating.

E- lr $28 These **results** can be correlated with those from the 5-arylmcthylhydantoins previously described 1,3 as shown in Figure 1.

E- $29 The superimposition of the hydantoin with the substrate (Figure 1a) placed the two oxidisable sites together and the phenyl group in a volume of space below the six-membered-ring plane.

E- ls $30 In a previous paper 3 we have suggested that enzymic oxidation of the benzyl group in the hydantoin leads to a benzylidene derivative which is susceptible to attack by an enzyme nucleophile as shown.

E- o (29) $31 When the spirocyclopropanobarbiturate structure was superimposed on that of the substrate (Figure I b), the hydrophobic substituent can be placed in a similar relative position to the phenyl group in the hydantoin inhibitors.

E- x $32 Attack on both activated systems by all enzymic nucleophile can then be envisaged as occurring from approximately the same direction.

E- lr $33 This **model** is consistent with all the results so far obtained in this series and, taken with the model obtained from regression analysis for the hydantoins, provides a basis for the design of more potent compounds with, for example, larger substituents on the benzene ring.

E- ve (9) $34 For the spirocyclopropanobarbiturates to have any value as inhibitors it is important to demonstrate that they have some selectivity of action.

a) P at / b) P $35 It could be argued that the cyclopropane ring, being doubly activated, would be sufficiently electrophilic to attack enzymes indiscriminately. 9

E- lr $35 The chemical reactivity of the spirohydantoins is certainly consistent with **this possibility.**

E- $36 It has been shown '° that the 5-spiro-(2-methyl) cyclopropanobarbiturate undergoes nucleophilic ring opening upon refluxing in solution in a range of alcohols (Scheme 3a).

E- lr $37 The location of the **substitution** was (surprisingly) found to be on the cyclopropane carbon bearing the substituent, the more hindered site.

E- lr $38 We have shown that **similar ring opening** takes place with the 2-phenylcyclopropyl derivative (Scheme 3b) and have confirmed the location of the substituent in the product by use of a deuterium-labelled derivative and by showing that no deuterium is lost (Scheme 3c).

E- lr $39 It is possible that the bond joining **this tertiary carbon atom** to the pyrimidine ring is weaker than the other possible cleavable bond because of unfavourable steric interactions between the ring and the substituent.

E- lr $40 Such **reactivity**, of course, provides a valuable insight into the probable mechanism by which these compounds act as enzyme inhibitors.

a) E- x b) E- ls $41 Although the possibility that the inhibitors (3) are not selective in their action has not been extensively investigated, we previously showed that compound (3a) was not an inhibitor of either horse liver alcohol dehydrogenase or of lactate dehydrogenase. 3

a) P at b) P2 $42 It was disappointing to find that none of the pyrazoles or spirocyclopropanohydantoins was found to be an irreversible inhibitor of DHODase.

E- x $43 However, both compounds (5b) and (7b) were weak competitive inhibitors whereas the parent unsubstituted compounds (5a) and (6) were totally inactive.

E-x \$44 The favourable interaction of aryl substituents with a hydrophobic pocket at the enzyme's active site is again indicated.

E- lr \$45 These **results** suggest that the enzyme has, not surprisingly, a preference for 6-membered-ring heterocycles and only when there are additional substituents present in 5-membered-ring compounds that can take part in ionic or adventitious hydrophobic bonding can the smaller ring form the basis for a significantly potent inhibitor.

E- or E- lr? \$46 The **importance of the hydrophobic group** was reinforced when the 5-arylidenebarbiturates (4) were tested as inhibitors.

E- lr \$47 These **compounds** were found to be the most reactive irreversible inhibitors so far investigated by us for DHODase.

E- ls \$48 As a result of their intense colour and the rapidity of the inhibition reaction ($t1/2 < 3$ min under standard conditions), accurate rate constants could not be obtained by the methods available.

E- \$49 Computer modelling of the electrostatic potential surfaces of the basic hydantoin and barbiturate skeletons in comparison with the substrate, using an algorithm developed by Dr. P. Bladon, suggests that the enzyme would recognise the barbiturates as substrate surrogates, but not recognise the hydantoins as such.

E- lr \$50 This **modelling procedure** automatically compares the potential surfaces of two compounds and obtains the best fit.

a) P at / b) P2 \$51 Our studies, together with the results obtained by others on peptidases,7 suggest that the structural limits within which substrate surrogates can be designed are likely to be narrow.

E- i \$52 This is not surprising in view of the demands inherent in the design concept expounded in the introduction.

E- i \$53 It suggests that a gradual distancing from the natural substrate is likely to be the best strategy in design, a concept that we are pursuing in our own studies of peptidases.

APPENDIX B5

Posture in structural chemistry article JMC (author *PL*).

(rephrasings are marked in bold. $ indicates the beginning of a sentence.)

TITLE
TI $1 Structural Studies on Tazobactam

ABSTRACT
TI $2 Tazobactam (3, CIoHI2N4O5S) is an effective inhibitor of bacterial §-lactamases.

E- i $3 It crystallizes with unit cell dimensions a = 10.230 (2) ~, ~ = 14.396 (2) , and c = 17.291(2) A in space group P212121.

E- ls $4 Compared to the related inhibitor sulbactam (2), which lacks the triazole ring, crystalline tazobactam exhibits very similar §-lactam geometry and the same S(1) envelope conformation of the thiazolidine ring.

a) E- x b) E- lr $5 However, in both independent molecules of 3 a triazole ring nitrogen atom accepts an intermolecular hydrogen bond; similar **interaction** by this moiety of 3 with a hydrogen-bond donor on the enzyme, which is impossible for 2, could account for its enhanced inhibitory power.

E- i $6 Semiempirical molecular orbital calculations show pronounced neative potential there.

E- $7 Molecular mechanics supports the hypothesis that the carboxyl group can rotate freely and the triazole ring can "flip".

INTRODUCTION

TI $8 The increasing threat posed by bacteria which have developed resistance to conventional chemotherapy as a consequence of §-lactamases has brought about the development of novel agents to fight this threat.

a) E- b) E-x $9 One avenue of research has resulted in a series of §-lactamase inhibitors which have no inherent antibacterial activity, but which can be administered in conjunction with long-established antibiotics to provide an effective means of treating infections with §-lactamase producing strains.

E- ls $10 Clavulanic acid~ (1) and sulbactam2 (2) have both been found to inhibit many clinically important §-lactamases and have been combined with §-lactamase-susceptible penicillins in clinical usage.3

E- x $11 More recently,4 the development of tazobactam (3, YTR-830, 3-methyl-7-oxo-3(lH-1 ,2,3-triazol-1-ylmethyl)-4-thia-1-azabicyclo[3.2.0]heptane-2-carboxylic acid, 4,4-dioxide) has resulted in a §-lactamase inhibitor with a very low toxicity, wide range of inhibition, and weak induction of §-lactamases.

E- ls $12 As part of our continuing studies on the interaction between §-lactamases and their inhibitors, we wish to report the crystal structure of tazobactam and to present initial molecular modeling studies on the possible conformations of this molecule.

METHODS AND PRELIMARY RESULTS

P - ts CRYSTAL STRUCTURE DETERMINATION

P2 $13 In order to facilitate computer-graphics modeling of this §-lactamase inhibitor, the crystal structure of 3 was determined.

E- $14 Crystallization from a 70:30 ethanol/water solution afforded colorless hexagonal prisms which were

found to contain two independent molecules of 3.

P (table) (13) $15 The structures (ORTEP6 drawing) with their numbering scheme are shown in Figure 1, and the relevant experimental data are summarized in Table I.

a) E- b) P- (al and table) $16 The principal differences between the primed and unprimed molecules are observed in the positions of the triazolylmethyl and carboxylic acid moieties, and the main torsion angle differences between the two molecules are given in Table II.

E- $17 The thiazolidine rings are puckered into an envelope conformation, with S(1) and S(1') out of the plane of the other four atoms by 0.813 (1) and 0.818 (1) i~, respectively, both with asymmetry parameters7 ~Cs of 1.3¡.

E- ls $18 The alternative penicillin conformation with S(1) in plane and the § carbon C(2) out of plane would have positioned the sulfone oxygen
atoms symmetrically.

a) E- x b) E-x $19 Nevertheless, the S(1) envelope conformation appears preferable since this is also observed in sulbactam8 (2) (~Cs = 6.1¡).

E- lr $20 This **same thiazolidine conformation** appears in X-ray and NMR studies on penicillin §-sulfoxides, where ~C~ = 5.3¡ for penamicillin9 and ~Cs = 2.5¡ for cloxacillinl¡ derivatives.

E- lr $21 It has been suggested9 that this **geometry** is inimical toward binding to the target transpeptidase enzyme; however, this may aid in binding to §-lactamases.

E- $22 An interesting feature apparent from the crystallographic study was the presence of intermolecular hydrogen bonding between the C(2)carboxylic acid proton and the N(4)triazolyl nitrogen.

E- ls $23 Examination of the Cambridge Crystallographic Database for compounds containing both triazole and carboxylic acid functionalities indicated literature precedent for this **type of hydrogen bonding**.ll l2

P (table) $24 Table III shows the appropriate hydrogen-bond lengths and angles.

P - ts Molecular Modeling Studies

P2 $25 The enhanced §-lactamase inhibitory activity5 of 3 compared to that of dimethyl analogue 2 is presumably due to the presence of the triazole function and thus we believe that this ring must participate in some favorable interaction within the active site of the §-lactamase.

E- x $26 In particular we wish to investigate the molecular basis of enzyme inhibition as postulated by Knowles.l3

E- i $27 This involves acylation of an active site serine by the §-lactam, followed by a nucleophilic attack on the C(3) position of the §-lactam by another active site residue.

E- lr $28 In order to study these **hypotheses**, it was necessary to establish the most favorableconformation of the molecule and to examine other possible conformations of the molecule which could be adopted.

E- ve (14) $29 The optimum geometries of both independent molecules of 3 in the crystal structure were determined with MoPAc,I4 employing the MNDO/ PM3 Hamiltonian.

E- lr $30 The **independent structures** optimized to virtually identical low-energy conformations, giving a final heat of formation of -76.8 kcal mol~l for the unprimed structure and -76.5 kcal mol-l for the primed structure.

E- overlay (29) $31 Before optimization, the corresponding figures were -38.0 and -32.8 kcal mol~l.

E- ve (17) $32 The thiazolidine ring was less puckered after optimization than in the crystal, and as a result, the triazole ring occupied a slightly different region of space.

E- ls $33 The C(1)S(1)-C(3)-C(4) torsion angles of 131.0 (2)¡ and 130.1(3)¡ for the unprimed and primed molecules decreased to 123.5¡ and 119.4¡ upon optimization.

E-x $34 The MOPAC calculations also suggest why 3 exhibits intermolecular hydrogen bonding between the C(4) triazole nitrogen and H(12) acid proton as described above.

E- x (refers to Table) $35 It can be seen that the areas of most negative potential (dark blue) are centered around the triazole ring nitrogens, and the carboxyl oxygens have a slightly less negative potential.

E- $36 The conformational flexibility of both independent molecules was investigated by using molecular mechanical calculations with the modified MM2 parametersl5 available within Chem-X.I6

E- ls $37 Separate rotation around the C(1) C(8) and C(8)-N(2) bonds at ca. 10¡ intervals followed by calculation of the molecular mechanics energy shows that for the unprimed molecule, the energy barrier to rotation about either of these bonds is no greater than 50 kcal mol-l.

E- ls $38 In the case of the primed molecule slightly more energy is required to "flip" the triazole ring through 180¡.

E- ve (37 $39 Figure 3, parts a and b, show plots of torsion angle against energy for rotation around the C(1)-C(8) and C(8)-N(2) bonds for both independent molecules.

E- ve (33) $40 The equivalent torsion angles for the molecules in the crystal structure are -77.0¡ and -100.1¡ for the unprimed molecule and -87.8¡ and -100.8¡ for the primed molecule.

P al (table) $41 Figure 4 shows the equivalent energy contour plot for the unprimed molecule.

E- x (refers to Table) $42 It can be seen that conformations of 3 with the triazole ring rotated through 180¡ occupy low-energy troughs, and it is possible that the energy required to overcome the barrier to rotation could be supplied to the molecules when in solution.

E- ls (E-lr?) $43 The only **other part of the molecule which is able to undergo free rotation** is the C(2) carboxylic acid group.

E- x $44 Conformational analysis of rotation around the C(2)-C(6) bond as described above indicated that the energy barrier to rotation is very low, with the difference between the lowest and highest energy conformations of both molecules no greater than 9 kcal mol-l.

E- x $45 It is therefore reasonable toC(1)-C(8)-N(2)-N(3) torsion angle (degrees) for unprimed and primed molecules assume that when in solution, sufficient energy can be obtained to rotate the carboxylic acid group into the most advantangeous conformation for binding to the enzyme.

P ts Comparison with Sulbactam.

P2 $46 In order to determine what effect, if any, the triazolylmethyl group exerts on the penam skeleton, the crystal structure of 3 was compared to that of the §-lactamase inhibitor sulbactam (2).

E- $47 Several parameters have been studied in an attempt to correlate the biological activity of §-lactams with structure.

E- lr $48 The **chemical reactivity** of the amide bond17 and a suitable distance between the p-lactam oxygen and the carboxylic acid carbon18 (the Cohen distance) have both been used in structure-activity relationships.

E- ls $49 The reactivity of the amide bond has been related to the C=O/CÑN bond lengths and the pyramidality of the §-lactam nitrogen.17

E- lr $50 These **parameters** were determined from the crystal structures of 25 and 3 and are summarized in Table IV.

E- i $51 It can be seen that values for the two independent molecules of 3 are very similar to those of 2 although the Cohen distance is at the upper limit of the range ascribed to active compounds19 (3.0-3.9 ~).

E- ve (49) $52 The pyramidality of the §-lactam nitrogen was expressed as the distance from this atom to the plane through C(2), C(3), and C(5).

E- lr $53 These **figures** indicate that the §-lactam N lies slightly further out of the plane of the three atoms surrounding it than the corresponding atom in 2.

E- x $54 Consequently, amide resonance is hindered and, as has been postulated by Woodward,17 the susceptibility of the §-lactam to nucleophilic attack should be slightly greater.

DISCUSSION

a) E- x b) E-x $55 It does not, therefore, appear that the triazolyl group exerts any major effect on the geometry of the §-lactam system of 3 (compared to 2), indicating that its enhanced §-lactamase inhibitory activity does not stem from any altered chemical reactivity.

E- overlay (51) or x $56 In general, the bond angles, bond distances, and torsion angles for the two independent molecules of 3 bear a very close similarity to the analogous data for 2.

E- lr $57 The **principal differences** were noted around the carboxylic acid functionality, with the C~H bond length decreased from 1.327 (6) A in 2 to 1.288 (4)1~ and 1.306 (5) ~ in the unprimed and primed molecules, respectively.

E- $58 Analysis of crystallographic data for 2 shows that an intermolecular hydrogen bond exists between the §-lactam carbonyl oxygen and the carboxylic acid proton.

E- lr $59 In the case of 3, the enhanced electron density present on the triazolyl ring makes possible the **hydrogen bonding described above**.

E- x $60 The possibility also exists for the formation of such a hydrogen bond between the triazole nitrogen and a suitable peptide residue within the enzyme active site.

E- lr $61 Work currently underway within these laboratories is directed toward identification of such **a residue** and examination of the binding of 3 to §-lactamases in an effort to rationalize its biological activity.

APPENDIX B6

Posture in chemistry article TL (authors *JE, JG*)

(rephrasings are marked in **bold**. Sentences are numbered by $)

TITLE
TI $1 Synthesis of Antiviral Nucleosides from Crotonaldehyde. Part 3.Total Synthesis of Didehydrodideoxythymidine (d4T)

Abstract

TI $2 The total synthesis of the antiviral agent d4T **3** from the epoxyalcohol **2**, itself derived from crotonaldehyde **1**, in 6 steps and 18% overall yield is described.

Introduction
TI $3 Inhibition of viral reverse transcriptase is currently the most established effective point of intervention for the treatment of retroviral diseases such as AIDS.

a) E- b) E-x $4 Modified 2'-deoxynucleosides lacking a 3'-hydroxyl group are often good inhibitors of HIV reverse transcriptase, and thus exhibit anti-HIV activity.

E- i $4 These include the currently approved therapies for AIDS or ARC, 3'-azido-3'-deoxythymidine (AZT),[3] and dideoxyinosine, ddI[4,5], and several other drugs currently in clinical trials as anti-AIDS drugs, including ddC.

E- x $5 The modified nucleoside didehydrodideoxythymidine, d4T, **3**, is also attracting current interest as another potentially clinically useful anti-HIV agent.[7]

E- ls $6 AZT insensitive HIV strains do not show cross resistance to d4T,[8] it readily crosses the blood brain barrier,[9] and its lower toxicity than, and comparable potency to, AZT suggest considerable potential for d4T as an anti-AIDS drug.[10]

a) E- / b) E- x $7 A number of syntheses from nucleoside starting materials,[11] or from carbohydrate derived materials such as ribonolactone,[12] have been reported, but no synthesis to date has commenced from non-chiral pool materials.

E- ls $8 As part of our program to develop novel and versatile synthetic routes to modified nucleosides,[13] we have recently reported the syntheses of the anti-AIDS drug AZT,[1] and of the anti-HIV agent ddC,[2] each in nine steps from the inexpensive achiral starting material, crotonaldehyde, **1**.

P- at Data analysis

P2 $9 Chirality was introduced by a Sharpless-Katsuki asymmetric epoxidation to give the common chiral epoxy alcohol, **2**.

E- ls (E-x?) $10 We now report that the epoxy alcohol, **2**, can also be elaborated in six steps to the anti-HIV agent, d4T (**3**), (Scheme I), and also to the 5'-acetyl-3'-thiophenylthymidine, **8a**, and 5'-acetyl-3'-selenophenylthymidine, **8b**, nucleoside analogues in four steps from **2**.

E- ls $11 Ring opening of the epoxy alcohol **2** with either thiophenol or selenophenol (1 equiv), was catalysed by diethylaluminum fluoride (other Lewis acids were ineffective), affording mixtures of the anticipated 1,2-diol products, **4**, (major product; 50% X = S; 30% X = Se), together with 8-10% of the undesired 1,3-diols, **6**, and 9-11% of the methyl glycosides, **5**.

E- lr \$12 The **minor products** were separated from the major diol by either flash chromatography or preparative tlc.

E- \$13 Cyclization of the diols **4** to the glycosides **5** proceeded in near quantitative yield, by employment of the conditions utilized previously in our syntheses of AZT[1] and ddC.[2]

a) E- lr b) E-x \$14 Combination of the glycosides (**5**) obtained from this **reaction** and from concomitant cyclization during the ring opening reaction, therefore affords 5a and 5b from epoxyalcohol **2** in overall yields of 48% and 29% respectively.

E- lr \$15 Acetylation of these **alcohols** to give the acetates 7 proceeded in ≥97% yield, after chromatography.

E- ls \$16 Vorbrüggen coupling[14] of these acetates with 2-3 equivalents of bis(trimethylsilyl)thymine catalysed by 2-3 equivalents of *t*-butyldimethylsilyl triflate (TBDMSOTf) in acetonitrile, yielded the 5'-acetyl-3'-thiophenylthymidine, (8a) and 5'-acetyl-3'-selenophenylthymidine, (8b), in 30% and 50% yields respectively, as anomeric mixtures.[12a, 12c]

E- \$17 Treatment of the seleno compound, 8b, with 1 equivalent of *m*-chloroperoxybenzoic acid (MCPBA) in dichloromethane at -5°C, and warming to room temperature over 2 hours, resulted in elimination of PhSeOH to give 5'-acetyl-d4T (9), in ≥95% yield.[15]

E- \$18 Deacetylation was effected in quantitative yields by treatment with methanolic ammonia to provide d4T, 3, and the a anomer, 10, as a 1:1.5 mixture.[16]

E- \$19 The ^{1}H NMR spectra of 9 and 3 matched those reported in the literature.[11a]

E- lr \$20 This **route** thus provides d4T in six steps and 18% overall yield from epoxyalcohol 2, and in 10 steps and 5% overall yield from crotonaldehyde.

E- ve (10) \$21 The 5'-acetyl-3'-thiophenyl-thymidine (8a) and 5'-acetyl-3'-selenophenylthymidine (8b), are obtained in overall yields of 16% and 19% respectively from epoxy alcohol 2.

E- lr (1 refocussing 2 reformulation) \$22 The completion of this **total synthesis**, together with those of AZT[1] and ddC,[2] clearly establishes this **methodology** as a general and versatile strategy towards the efficient synthesis of a range of important antiviral modified nucleosides from cheap achiral starting materials.

E- lr \$23 Further work on the extension of this **methodology** to other important types of modified nucleosides is underway.[17]

APPENDIX B7

Posture in cancer research article BJ (authors *HM and MT*).

(rephrasings are marked in bold. Sentences are numbered by $)

TITLE
TI $1 Metabolic substrate utilization by tumour and host tissues in cancer cachexia.

ABSTRACT
TI $2 Utilization of metabolic substrates in tumour and host tissues was determined in the presence or absence of two colonic tumours, the MAC16, which is capable of inducing cachexia in recipient animals, and the MAC13, which is of the same histological type, but without the effect on host body composition.

E- ls $3 Glucose utilization by different tissues was determined in vivo by the 2-deoxyglucose tracer technique.

E- ls $4 Glucose utilization by the MAC13 tumour was significantly higher than by the MAC16 tumour, and in animals bearing tumours of either type the tumour was the second major consumer of glucose after the brain.

E- lr $5 This **extra demand for glucose** was accompanied by a marked decrease in glucose utilization by the epididymal fat-pads, testes, colon, spleen, kidney and, in particular, the brain, in tumour-bearing animals irrespective of cachexia.

E- ls $6 The decrease in glucose consumption by the brain was at least as high as the metabolic demand by the tumour.

E- i $7 This suggests that the tissues of tumour-bearing animals adapt to use substrates other than glucose and that alterations in glucose utilization are not responsible for the cachexia.

E- $8 Studies in vitro showed that brain metabolism in the tumour bearing state was maintained by an increased use of lactate and 3 hydroxybutyrate, accompanied by a 50% increase in 3-oxoacid CoA-transferase.

E- i $9 This was supported by studies in vivo which showed an increased metabolism of 3hydroxybutyrate in tumour-bearing animals.

E- x $10 Thus ketone bodies may be utilized as a metabolic fuel during the cancer-bearing state, even though the nutritional conditions mimic the fed state.

INTRODUCTION
TI $11 Biochemical changes in host tissues frequently occur in cancer patients, and depletion of host adipose tissue and muscle protein in cancer patients is an important parameter determining overall survival [1].

E- ls $12 The reason for depletion of host tissues is not known, but is thought to arise from differences in metabolism in the tumour-bearing state.

E- x $13 Thus some studies have shown that the daily energy expenditure and the resting metabolic rate are much higher in cancer patients than in controls, even though the energy intake was not significantly different between the two groups [2].

a) P at b) P2 $14 Although anorexia is common in cancer patients [3], numerous studies have documented abnormalities of host metabolism, which together indicate that cancer cachexia is not the same as simple starvation [4].

a) E- ve (12) b) E-x (overides E- lr) $15 There are marked alterations in the carbohydrate metabolism of host tissues of both animals and patients with cancer, but how these **alterations** contribute

to the cachexia remains speculative.

E- $16 Glucose metabolism in cancer cells is known to be elevated, owing to an altered membrane transport, which leads to an increased intracellular concentration [5], with consequent increased lactate production by some tumours [6].

E- i $17 This in turn leads to an increased conversion of lactate into glucose in the liver via the Cori cycle, the activity of which is also found to be elevated in cancer patients with progressive weight loss, showing that lactate production rates are higher in these patients [7].

E- ls $18 Decreased blood glucose values are often found in tumour-bearing animals [8], suggesting that the glucose availability is not sufficient to provide for both host and tumour.

E- x $19 In addition, a decreased plasma insulin level is often found [9], together with an elevated rate of gluconeogenesis from non-carbohydrate precursors [10].

a) P at b) P2 $20 To try to determine the biochemical changes responsible for, or accompanying, cachexia, we have employed an experimental murine colon adenocarcinoma (MAC16), which produces extensive loss of host body weight, accompanied by decreases in both fat and lean body mass, but without a fall in food intake [11,12].

E- lr $21 This **tumour** is useful for the study of the mechanisms of cachexia, since in some cases growth of the tumour is not accompanied by weight loss [13], suggesting that the presence of the tumour alone is insufficient to produce the effects on host body tissues.

E- x $22 In addition, tumours of the same histological type are available which grow without an accompanying cachexia (e.g. MAC 13).

E- ve (18) $23 This study measures changes in glucose utilization in host and tumour tissues in animals with or without cachexia, in order to elucidate if changes in carbohydrate metabolism are related to the cachectic state.

MATERIALS AND METHODS

TI $24 Pure-strain NMRI mice were bred in our own colony and were fed on a rat and mousc breeding diet (Pilsbury, Birmingham, U.K.) and water ad libitum.

E- ls $25 Fragments of either the MAC16 or MAC13 tumour were implanted into the flank of male NMRI mice (starting weight 24 26 g) by means of a trocar, as described in [11,12].

E- ls $26 Animals bearing the MAC16 tumour develop weight loss 10 12 days after the transplantation (average tumour weight 200 mg), and, when weight loss developed, the animals were regarded as cachectic (average weight loss 2 4 g).

E- ls $27 Occasionally tumours grew without the development of weight loss, and these animals were regarded as non-cachectic (average tumour weight 200 mg; about 5% of the total transplanted).

E- ls $28 Animals bearing the MAC13 tumour were used 10 -12 days after tumour transplantation, when the tumour became palpable (average tumour weight 200 mg).

{TABLE}$29 The daily food intake per mouse [63.2+2.5 kJ 2DG, 2-deoxyglucose; 2DGP, 2-deoxyglucose 6-phosphate; [:1H]2DG, 2-deoxy-D-[2,6-3H]glucose; [l4C]2DG, 2-deoxy-D[I -l4C]glucose (15.1+0.6 kcal)] in animals bearing the MAC16 tumour did not differ from that of non-tumour-bearing controls [64.0 + 1.25 kJ (15.3+0.3 kcal)],whereas in animals bearing the MAC13 tumour the daily food intake [68.6+1.25 kJ (16.4+0.3 kcal)] was significantly increased ($P < 0.01$).

P ts $29 GLUCOSE UTLIZATION.

P2 $30 The extent of glucose utilization by different tissues was investigated by using the 2-deoxyglucose tracer technique [14].

E- $31 Briefly, animals were starved overnight and throughout the experiment, but given water ad libitum.

E- i $32 The following morning they were injected intravenously with 30 ,uCi of 2-deoxy-D-[2,63H]glucose ([3H]2DG; sp. radioactivity 42mCi/mmol; Amersham International, Amersham, Bucks., U.K.)/kg body wt. in 200 1/1 of normal saline.

E- lr $33 To determine the retention of 2deoxyglucose 6-phosphate (2DGP) by different tissues, a **second intravenous injection** of 3 ,uCi of 2-deoxy-D-[1-l4C]glucose (['4C]2DG; sp. radioactivity 56 mCi/mmol; Amersham International)/kg was administered 35 min after the injection of the [3H]glucose.

E- ls $34 Blood was removed from animals by cardiac puncture, while under anaesthesia, at specified time intervals, and the decay of radioactivity in the blood was monitored for 60 min [14].

E- ve (30) $35 Blood glucose concentration was determined on whole blood by using the o-toluidine reagent kit (Sigma Chemical Co., Poole, Dorset, U.K.).

E- ls $36 The concentration of radioactivity in the blood was determined on deproteinized neutralized samples with a dual 3H/I4C analyser.

E- $37 The accumulation of phosphorylated metabolites of 2DG was measured in selected tissues at the 60min time point [14,15].

E- ve (30) $38 Glucose utilization was calculated from the equation [14]: {R Cm * (7~LC 1' ' dt} where R is tissue glucose metabolic rate (nmol/min per g), C ,* is the concentration of phosphorylated metabolites of 2DG in the tissue (d.p.m./g) at t = 60 min, C" is the blood glucose (nmol/ml), Cl)* is the concentration of [3H]2DG in the blood (d.p.m./ml) and LC (lumped constant) is a dimensionless correction factor for discrimination against 2DG in glucose metabolic pathways and was determined by the method of Ferré et al. [16].

E- ls $39 The value determined for various tissues, 0.46 [17], corresponds favourably to that previously reported [14 16].

P ts $40 GLUCOSE UTILIZATION IN VITRO BY TUMOUR CELL LINES.
P2 $41 Both the MAC16 and MAC13 cell lines were derived rrom the corresponding solid tumours and were maintained in vitro in RPMI 1640 culture medium containing 10o,') (v/v) foetal-calf serum (Gibco Europe, Paisley, Scotland, U.K.) under an atmosphere of 5 "(~ CO2 in air.

E- $42 For measurement of the rate of CO2 production, the medium, containing cells at a density of 4 x 104 ml 1, was supplemented with 0.2 IICi of D-[U-I4C]glucose (sp. radioactivity 270 mCi/mmol; Amersham International ml in a flask with a centre well and sealed with a rubber seal.

E- ls $43 At specified time points 0.3 ml of 0.3 M-NaOH was injected into the centre well, 0.3 ml of 5 M-HC104 was added to the medium, and the flasks were left overnight.

E- ls $44 The radioactivity in the alkali was determined with a Packard Tri-carb 2000 CA liquid-scintillation analyser, and the cell number was enumerated with a Coulter electronic particle counter.

E- ve (38) $45 For determination of glucose consumption and lactate production, cells were resuspendcd at a density of 4 x 104 ml 1, and at specified time intervals medium was removed and the concentration of glucose was determined with the o-toluidine reagent kit, and the concentration of lactate by the method of Gutmann & Wahlefield [18].

P ts $46 FUEL UTILIZATION BY LIVER, BRAIN AND TUMOUR IN VITRO
P2 $47 Slices of liver and brain from control and tumour-bearing animals, together with tumour slices, were incubated individually in a Krebs Ringer bicarbonate solution containing glucose (2.2 mM), DL-3-

hydroxybutyrate (0.1 mM) and lactate (8 mM).

E- lr \$48 In each **fuel mixture** only one of the three substrates was radioactive.

E- \$49 The radiolabelled compounds D-[U-I4C] glucose (sp. radioactivity 273 mCi/mmol) and L-[U-I4C]lactic acid (sp. radioactivity 169 mCi/mmol) were purchased from Amersham International D-(~ -hydroxy[3-l4C]butyric acid (potassium salt; sp. radioactivity 44.3 mCi/mmol) was purchased from New England Nuclear, Stevenage, Herts., U.K.

E- ve (47) \$50 The slices were incubated with 10 ml of fuel mixture containing 2.5 ,uCi of the labelled fuel in a flask containing a centre well and sealed with a Suba-seal.

E- \$51 Samples were gassed with 02/C02 (19:1) and incubated at 37 °C for I h.

E- x \$52 At this time 0.3 ml of 5 M-HC104 was added to the medium to liberate the CO2, which was collected in 0.3 ml of 0.3 M-NaOH added to the centre well.

E- \$53 After standing for 4 h to allow complete absorption, the contents of the centre well were added to 10 ml of Optiphase Hi-safe 3 scintillation fluid and the radioactivity was determined.

P at \$54 Determination of brain 3-oxoacid CoA-transferase.

a) P b) E- ve (47) \$55 Brains were homogenized in 10 mM-Tris, pH 7.4, containing 0.25 M-sucrose and I mM-2-mercaptoethanol, and enzyme activity was determined in the supernatant fraction prepared by centrifugation at 30000g for 20min.

E- \$56 The rate of acetoacetylCoA formation from 0.1 mM-succinyl-CoA was determined by the method of Williamson et al. [19].

P ts \$57 Effect of tumour-bearing state on 14CO2 production from 3-hydroxybutyric acid

P \$58 Male NMRI mice bearing either the MAC16 tumour with weight loss or the MAC13 tumour, or non-tumour-bearing controls, were injected intravenously with 5OIICi of D-(—)-1~hydroxy[3-l4C]butyric acid (sp. radioactivity 44.3 mCi/mmol) (Amersham International)/kg and were placed in airtight metabolic cages with the entry air being pumped through CaCO.~ (solid) to absorb any CO2.

E- ve (53) \$59 Metabolically produced 14CO2 was trapped in ethanolamine/ethoxyethanol (1:4, v/v), and samples were taken at specified time intervals and the radioactivity was determined directly in Optiphase scintillation fluid (FSA Laboratory Supplies, Loughborough, Leics., U.K.).

RESULTS

P ts \$60 In order to understand the glucose dynamics of the MAC16 and MAC13 tumours in vivo, and the effects on host organs with the development of cachexia, glucose utilization was investigated by the 2DG tracer method [14 16].

a) P b) E- x \$61 Blood glucose levels in animals bearing the MACl 3 tumour that had been starved overnight were similar to those in control non-tumour-bearing animals over a 60 min time period (Fig. la), whereas animals bearing the MAC16 tumour were hypoglycaemic as compared with either group, irrespective of the development of cachexia, although this was only significant at the 60 min time point.

E- x \$61 However, the rate of disappearance of the label from [:IH]2DG (Fig. 1~7) or from [l4C]2DG (Fig. Ie) did not differ between the four groups.

E- ve (62) \$62 The tissue glucose metabolic rate (R~) of control mice and animals bearing either the MAC13 tumour or the MAC16 tumour, with or without cachexia is shown in Table 1.

E- ve (60) $63 Glucose utilization by the MAC13 tumour was significantly ($P < 0.05$) greater than that by the MAC16 tumour, probably owing to the higher rate of cell replication (doubling time of MAC13 tumour 7 days; doubling time of MAC16 tumour 10-12 days).

E- ls $64 There was no difference in glucose utilization by the MACl 6 tumour in the presence or absence of cachexia (Table 1).

E- ve (62) $65 The R values for testes, colon, spleen, kidney and particularly brain ($P < 0.005$) were significantly decreased in animals bearing the MAC16 tumour.

E- ls $66 The large decrease in the Rg value for brain was not specific for the cachectic state, since it was also seen in animals bearing the MAC16 tumour in the absence of cachexia, and in MAC13-tumour-bearing animals.

E- ve (63) $67 The contribution of the various organs to glucose utilization in tumour-bearing animals is shown in Table 2.

E- ls $68 The magnitude of the contribution to glucose utilization by the various organs depends on both the R~ value and the weight of a particular organ.

E- $69 In all tumour-bearing animals the tumour was the second major consumer of glucose after the brain.

E- i $70 This placed a high demand on the host, and was accompanied by a marked decrease in glucose utilization by the brain.

a) E- lr b) E- x $71 In animals bearing the MAC13 tumour, **glucose consumption by the brain** decreased by an amount equal to that used by the tumour, whereas, in animals bearing the MAC16 tumour, brain glucose consumption was decreased to a greater extent than glucose consumption by the tumour.

E- ls $72 In animals bearing the MAC16 tumour, glucose consumption was also significantly decreased in epididymal fatpads, testes, colon, spleen and kidney, but was not related to the presence of cachexia.

a) P at b) P2 $73 To investigate whether cachexia altered the retention of 2DGP in the various organs, a sequential double-labelling technique was applied, followed by an analysis of the two labels in 2DGP.

E- ls $74 Since there was a marked initial decay of the precursor in the blood, the bulk of the [3H]2DGP was synthesized in the tissues during the initial 35 min of the labelling period, and the 3H/I4C ratio of tissue 2DGP was measured at the end of the experiment.

E- ls $75 Loss of 2DGP from the tissue would affect the 3H component of the ratio more than the 14C component, and therefore the iH/I1C ratio of 2DGP in the tissues was a measure of the retention of 2DGP, i.e. a low ratio indicated a high rate of loss.

E- ls $76 Since 10 times as much 3H radioactivity was administered as 14C, the :iH/I4C ratio would be expected to be near 10 if both radioisotopes were incorporated into 2DGP at the 60 min time period and there was no loss of DGP.

E- i $77 This was true for most control tissues, except for lung, colon, brain and kidney, in which the ratios were much lower (Table 3).

E-lr $78 The lower ratio for brain has previously been reported [14], and arises from an increased rate of loss of 2DGP from this **tissue**.

E- ls $79 When compared with control animals and MAC16-tumour-bearing animals without cachexia, the 3H/I4C ratio for testes, lung and brain was significantly higher in cachectic animals bearing the MAC16 tumour (Table 3).

E- i $80 This suggests a slow incorporation of 2DG into 2DGP, a low rate of utilization and, assuming that the energy requirements remain constant, that the tissues of cachectic animals adapt to use metabolic substrates other than glucose.

E- ve (66) $81 The decrease in glucose utilization by organs in tumour bearing mice must be accompanied by an alteration in the preference of the substrate for energy production.

a) P at b) P2 $82 To determine the preferred metabolic substrate, the utilization of glucose, lactate and 3-hydroxybutyrate was determined by using a substrate mixture with a high lactate concentration.

P (refers to Table) $83The results for liver, brain and tumour are shown in Table 4.

E- ve (82) $84 At the concentration employed, lactate was the most important oxidative substrate for the liver in control and tumour-bearing animals, with the rate of utilization by liver slices from animals bearing the MAC13 tumour being significantly ($P < 0.005$) elevated over that of non tumour-bearing controls (Table 4).

E- $85 Both glucose and 3-hydroxybutyrate consumptions by the liver were low, and, although the former did not change in the tumour-bearing state, consumption of 3-hydroxybutyrate was significantly higher ($P < 0.001$) in animals bearing the MAC13 tumour.

E- ls $86 Lactate was an important metabolic fuel for the tumour, and the rate of utilization was significantly greater ($P < 0.001$) in the MAC13 tumour.

E- x $87 Although the glucose consumption rate was not significantly different for the two tumour types, the rate of utilization of 3-hydroxybutyrate was much higher in the MAC 13 tumour.

E- ls $88 For brain from non-tumour-bearing animals, glucose was the predominant metabolic fuel, followed by lactate, possibly owing to the supra-physiological lactate concentration, whereas in the brain of animals bearing the MAC16 tumour with cachexia, the rate of glucose utilization was decreased by two-thirds ($P < 0.05$) and was also significantly less than that in brains of animals bearing the MAC13 tumour ($P < 0.005$).

E- lr $89 This **metabolic deficit produced by a decreased glucose consumption** was accompanied by a marked increase in both lactate ($P < 0.05$) and 3hydroxybutyrate ($P < 0.01$) utilization in the brains of animals bearing the MAC16 tumour, when compared with non-tumour bearing controls.

E- lr $90 This **increased 3-hydroxybutyrate utilization in the brains of tumour-bearing animals** is probably due to an increase in the level of 3-oxoacid CoA-transferase (EC 2.8.3.5), the rate-limiting enzyme in ketone-body utilization.

E- x $91 Thus animals bearing both the MAC 16 and MAC 13 tumours show a significant ($P < 0.05$) elevation of brain 3-oxoacid CoA-transferase activity (I . 8 + 0. 2 and I . 5 `+ 0 . I 1I mol/min per mg of protein respectively) as compared with non-tumour-bearing controls (1.2+ 0.1 / l mol / min per mg of protein).

E- $92 Production of '4CO2 after intravenous administration of D(—)3-hydroxy[3-l4C]butyrate was rapid in all groups of animals and was significantly ($P < 0.05$) greater in animals bearing both the MAC16 and MAC13 tumours (Fig. 2).

E- i $93 This suggests that in tumour-bearing animals brain metabolism is supported by an increased utilization of 3-hydroxybutyrate, as occurs in starvation, and this appears to be independent of the development of cachexia.

DISCUSSION

TI $94 Glucose is an important metabolic substrate, particularly for solid tumours [20], possibly owing to a poor vascularization, which would militate against oxidative metabolism.

E- ls $95 In a study of energy production by a rapidly growing hepatoma cell line, it was estimated that about 60% of the total ATP was derived from glycolysis and 40~/O from oxidative phosphorylation [21].

E- ve (91) $96 In the present study of the glucose consumption by the colon adenocarcinomas, the MAC series has been shown to be high (0.7-1.1 ,umol/min per g) and comparable with that of the brain (0.8 llmol/min per g).

E- x $97 Interestingly, the rate of glucose utilization by the MAC16 tumour was lower than that by the MAC13 tumour and independent of the production of cachexia in the animal, suggesting that the high glucose consumption by the tumour is not sufficient to explain the loss of host body compartments.

E- ls $98 The tumour-bearing state also induces profound changes in glucose utilization by host organs, and in particular by the brain, where glucose utilization is severely depressed.

E- lr or **E- ve (98)** $99 Previous studies [22] have shown that **glucose consumption** by murine peritonealexudate macrophages is suppressed by pleural effusions, ascites fluids and sera from patients with advanced primary lung and gastric cancers.

E- $100 Inhibition of glycolysis was ascribed to an inhibition of the rate-limiting key enzyme in the glycolysis pathway, D-fructose-6-phosphate l-phosphotransferase, although the nature of the inhibitory factor is not known.

E- ve (98) $101 Certainly, suppression of glucose utilization by host organs would be advantageous to the tumour, in view of the high glucose utilization.

a) E-x or E- ve (97) b) E-x $102 A decreased glucose uptake in the peripheral tissues of cancer patients has also been reported [23], possibly owing to a decreased insulin response, although this would not explain the decreased glucose utilization by the brain.

E- overlay (102) $103 The decreased glucose utilization by peripheral tissues in the tumour-bearing state suggests an increased dependence on fat as an energy source.

E- ls $104 Certainly, in patients with malignant cachexia, a greater proportion of oxidative metabolism appears to arise from fatty acids, particularly when exogenous glucose is available [24].

E- ls $105 The inability of glucose loading to suppress oxidation of fatty acids in cachectic patients suggests that there is a disturbance in normal homoeostatic mechanisms involving the utilization of glucose and other important fuel sources.

E- $106 We have recently observed an increased lipogenesis from glucose in host tissues in animals bearing both the MAC16 and MAC13 tumours (H. D. Mulligan & M. J. Tisdale, unpublished work).

E- i $107 This suggests that glucose is diverted away from host tissues, not only for direct utilization by the tumour, but also as an important starting material for the synthesis of substances that may be required for tumour growth.

E- x $108 In addition, if some glucose is not oxidized directly, but is first converted into fat before supplying utilizable energy, there is a significant energy cost to the organism, which could lead to loss of body weight if energy intake is not increased.

E- $109 In order to determine the predominant metabolic fuel used by tumour and host tissues, we have employed an assay system in vitro using slices of brain, liver and tumour fed with a fuel mixture of glucose (2.2 mM), D-3-hydroxybutyrate (0.05 mM) and lactate (8 mM) [25].

a) E- lr b) P-at c) P $110 This **system** paralleled the study in vivo using 2DG, in that, although glucose was the predominant metabolic fuel for the brain in non-tumour-bearing animals, its importance in oxidative metabolism was decreased in the tumour-bearing situation, especially in animals bearing the cachexia-inducing MAC16 tumour, where its use was replaced by lactate and 3hydroxybutyrate.

E- ls $111 Ketone bodies have been shown to replace glucose as the predominant fuel for brain metabolism during starvation [26].

a) E- x b) E-x $112 Thus the brains of cachectic animals bearing the MAC16 tumour resemble those during starvation, although the food intake is not decreased [11,12].

E- ve (111) $113 In the rat, the major factor influencing the rate of utilization of ketone bodies has been suggested to be the concentration in the blood [19].

a) E- x b) E- ls $114 However, although extensive mobilization of adipose tissue is observed in animals bearing the MAC16 tumour, no elevations in plasma or urinary levels of ketone bodies have been detected [II].

a) E- x b) E- ls $115 In addition, in the situation in vitro, the concentrations of glucose, lactate and 3-hydroxybutyrate in the incubation medium remained constant, suggesting that the enzymes which regulate ketone-body utilization in the brain of tumour-bearing animals may be altered.

E- l r $116 This **hypothesis** was confirmed by measurements of the levels of 3-oxoacid CoA-transferase, the rate-limiting enzyme in ketone-body utilization, which was shown to be significantly enhanced in animals bearing both the MAC]6 and MAC 13 tumours.

E- x $117 In contrast, the level of acetoacetyl-CoA thiolase was not significantly different between tumour-bearing (18 .3 + 4.1 ,umol/min per mg of protein) and non-tumour-bearing (22.0 + 6.6 ,umol/min per mg of protein) animals.

E- x $118 We have previously shown that 3-oxoacid CoA-transferase is also present in the MAC16 tumour, but at decreased Ievels compared with normal colon [27].

E- ve (105) $119 The activity of the enzyme in liver is low or non-detectable, and coincides with a low rate of utilization of 3-hydroxybutyrate by liver slices in vitro, whereas the rate of oxidation by tumour tissue was 3 10 times higher.

E- ve (86) $120 For the MAC13 tumour, lactate was the most important metabolic fuel at the high concentration employed, and lactate has also been shown to be utilized by a number of rat tumours in vivo [28].

E- ls $121 Studies in vitro with the two tumour cell lines showed that, although the total glucose consumption did not differ, oxidation to CO2 was greater in MAC13 cells ($P < 0.01$), whereas lactate production was higher in MAC16 cells ($P < 0.001$).

E-i $122 This suggests that there may be some difficulty in the oxidation of reduced nicotinamide nucleotides by the MAC16 tumour.

E- x $123 Thus the high metabolic demand for glucose placed on the host in the tumour-bearing state is accompanied by a decreased utilization by host organs, principally the brain, accompanied by utilization of ketone bodies as an alternative metabolic fuel.

E- ls $124 Utilization of ketone bodies appears to result from an increased activity of 3-oxoacid CoA-transferase without an alteration in substrate availability.

a) E- lr b) E- x $125 This **change** appears not to be specific for the cachectic state, although the increased metabolic demand on the host is associated with an increased food intake in animals bearing the MAC 13 tumour, whereas animals bearing the MAC 16 tumour have the same food intake as non-tumour bearing controls.

E- x $126 Thus cachexia could arise from an inability of the host to adapt to the increased metabolic requirements of the tumour-bearing state.

APPENDIX B8

Posture in cancer research article JNCI (author *MT*).

(rephrasings are marked in **bold**. Sentences are numbered by $)

TITLE

TI $1 Lipolytic Factors Associated With Murine and Human Cancer Cachexia.

ABSTRACT

TI $2 We have identified a lipolytic factor in extracts of a cachexia-inducing murine carcinoma (MAC16) that shows characteristics of an acidic peptide and appears to be composed of three fractions of apparent molecular weights corresponding to 3 kd, 1.5 kd, and 0.7 kd, as determined by exclusion chromatography.

E ls $3 Material with identical chromatographic and molecular weight characteristics was also present in the serum of patients with clinical cancer cachexia but absent from normal serum, even under conditions of starvation.

E ve ($2) $4 The MAC16 lipid factor, when injected into animals bearing the non-cachexia-inducing tumor MAC13, was capable of inducing weight loss without a significant reduction in food intake.

E-x (overrides E- lr) $5 Similar **lipolytic material**, although in lower concentration, was also found in the MAC13 tumor extracts.

E lr $6 These **findings** suggest that cachexia may arise from the enhanced expression of a lipolytic factor associated with tumor cells.

E- $7 Recently, considerable attention has been directed toward the isolation and identification of the factors responsible for the complex metabolic changes associated with cancer cachexia.

E-lr $8 Forefront in this **role** is tumor necrosis factor alpha (TNF-a), which has been shown to be homologous to cachectin (1), a macrophage product that mediates cachexia in animals infected with trypanosomes (2).

E ls $9 Tumor cells transfected with the human TNF-(cachectin) gene produce significant weight loss and severe cachexia in recipient animals (3).

E x $10 Weight loss associated with TNF-a, however, is always accompanied by marked anorexia (3,4), although loss of both muscle and adipose tissue in both rats and humans frequently precedes a fall in food intake (5).

a) E x b) E-ls $11 Although some studies have shown an elevated serum level of TNF-a in cancer patients (6) and in children with malignancies (7), other studies report undetectable plasma levels of endogenous TNF-a in cancer patients (8,9), including patients with cancer cachexia.

E x $12 Thus, the relationship between the serum TNF-a level and the development of cancer cachexia in humans is weak, since even the studies reporting elevated serum TNF-a showed no correlation with weight loss (6,7).

a) E x b) E x $13 A number of clinical trials of recombinant TNF-a have also been reported, and in none was any clinical evidence of accelerated cachexia demonstrated, although in some, anorexia was present during administration (10).

E lr $14 These results suggest that other factors may be the mediators of cachexia in cancer patients.

E ls $15 As a model of cancer cachexia, we have utilized a murine colon adenocarcinoma (MAC 16) solid tumor transplant that induces weight loss at small tumor burdens and without a reduction in food and water intake (4,11).

E ls $16 We have been unable to detect elevated levels of TNF-a in either tumor extracts or serum of animals bearing the MAC16-induced tumor, even in response to endotoxin (4).

E x $17 Weight loss, however, is associated with the presence, in both the tumor and the serum of cachectic animals, of a catabolic factor that directly induces lipolysis in adipose tissue and thus has the characteristic of a cachectic factor (12).

E ls $18 Both weight loss in vivo and lipolysis in vitro are suppressible by insulin and 3-hydroxybutyrate, both of which are also effective inhibitors of cachexia in vivo (13,14).

E x $19 We have attempted to characterise further the lipolytic factor associated with the MAC16 tumor and to extend these observations to cachectic human cancer patients.

Materials and Methods

TI $20 Pure-strain NMRI mice bred in our own colony (Bantin and Kingman Laboratory, Hull, England) were fed a rat and mouse-breeding diet (Pillsbury, Birmingham, England) and water ad libitum.

E-ls $21 Fragments of the MAC16 or MAC13 tumor, excised from donor animals, were implanted into the flank of male NMRI mice by means of a trocar, as previously described (11,12).

E ls
$22 Tumors were removed from animals with established weight loss, homogenized (10% wt/vol) at 4¡C in Krebs-Ringer bicarbonate buffer (pH 7.6), and centrifuged for 10 minutes at 3000g to remove debris.

E
$23 The supernatant was used for further characterization.

E $24 Blood was removed from both normal subjects and cachectic cancer patients, allowed to clot at room temperature (approximately 10 minutes), and centrifuged immediately. The serum was then separated and stored atÑ70¡C until needed.

P att Chromatographic Characterization

P 2 $25 The freshly prepared crude tumor extract or serum samples were fractionated by anion-exchange chromatography by use of a DEAE-cellulose column with elution under a salt gradient.

E ls $26 The tumor supernatant (containing 1.3 mgof protein) or serum samples (1 mL) were applied to a DEAE-cellulose column (dimensions: 1.6 cm X 30 em) equilibrated with 10 mM phosphate (pH 8.0).

E $27 Active material was eluted from the column by use of a linear gradient of 0.08 to O.2MNaCI in 10 mM phosphate (pH 8.0).

E ls $28 The column was eluted at a flow rate of 30 mL/hour, and the effluent was assayed for lipolytic activity with the use of epididymal adipocytes obtained from BALB/c or MFI mice, as previously described (Il).

E ve (27) $29 Samples contained 105 to 2 X 105 adipocytes per milliliter, and the incubation temperature was 37¡C for 2 hours.

E $30 The concentration of glycerol released was determined enzymatically by the method of Wieland (15).

E ls $31 Control samples containing adipocytes alone were analyzed to determine the spontaneous glycerol release.

E ls $32 Lipolytic activity was expressed as micromoles of glycerol released per 105 adipocytes per 2 hours.

E ls$33 Effluent fractions from the DEAE-cellulose column that possessed significant lipolytic activity were concentrated by vacuum dialysis, and the concentrate was applied to a Sephadex G50 column (1.6 em X 30 em) equilibrated with 10 mMphosphate (pH 8.0) and eluted at a flow rate of 15 mL/ hour.

E overlay(32) $34 The effluent was collected in l-mL fractions, and the lipolytic activity was determined as above.

P at $35 Effect of Tumor Lipolytic Factor on Body Weight of Animals Bearing the MAC13 Tumor

P2 $36 Active material was obtained from MAC16 cells growing in RPMI-164G medium with 10% fetal calf serum under an atmosphere of 5% CO2 in air, while an equal volume of medium from L1210 leukemia cells was purified by the same procedure and used as a control.

E $37 Medium was fractionated by Sephadex G 150, Biogel P4, and hydrophobic chromatography by use of a C18 column (16).

a) E (x?) b) E-lr $38 We did not attempt to assess the in vivo catabolic activity of the individual fractions, and we used a mixture for this **study**.

E ve (19) $39 Female NMRI mice bearing the MAC13 adenocarcinoma (starting wt, 18 to 20 g) were administered 50 IIL of the lipolytic factor (corresponding to 0.05 llmol of glycerol released per 105 adipocytes per 2 hours) or 50 IIL of the fractionated L1210 leukemia cell medium (purified by the same procedure but which did not possess any lipolytic activity) twice daily by intraperitoneal injection, and body weight and food intake were determined.

RESULTS

E ve (27) $40 The lipolytic activity from the MAC 16 tumor has been further characterized by DEAE-cellulose chromatography of tumor extracts (Fig l,A).

E ve (28,30) $41 Active material was retained on the column and could be eluted as four successive peaks under the influence of a linear salt gradient (0.08 to 0.2 M NaCl).

E lr $42 Each peak of lipolytic activity from the DEAE-cellulose column, when subjected to Sephadex G50 exclusion chromatography, gave three fractions of distinct apparent molecular weights corresponding to 3.0 kd, 1.5 kd, and 0.7 kd (Fig l,B).

E ls $43 When subjected to DEAE-cellulose chromatography, serum from cancer patients with cancer cachexia showed a large peak of nonretained activity and an activity-distribution pattern of a form similar to that of the MAC 16 tumor when it was subjected to a salt gradient (Figs 2,A and 2,B).

E x $44 Moreover, control samples of normal human serum, although containing the nonretained activity, contained no corresponding fractions of retained lipolytic activity (Fig 2,C).

E E-x b) E-x $45 Exclusion chromatography, with use of Sephadex G50 and when applied to each of the peaks of activity eluted under the influence of a salt gradient from the serum of the cachectic cancer patients, again gave three fractions of apparent molecular weights corresponding to 3.0 kd, 1.5 kd, and Q7 kd equal to those for the MAC 16 tumor (Flg 3,A), whereas serum from control subjects contained no corresponding peaks of activity, even after 24 hours of starvation (Fig 3,B).

a) E b) E-x $46 As previously reported (12), extracts of the MAC13 tumor, which does not produce cachexia in recipient animals, were capable of causing an enhanced lipolysis in murine adipocytes, although

to a much lower extent than extracts of the MAC 16 tumor.

a) P-at b) P2 \$47 To investigate whether the material responsible for lipolysis in extracts ofthe MAC13 tumor was similar to that found in the MAC16 tumor, we prepared a more concentrated tumor extract and applied it to a DEAEcellulose column (Fig 4,A).

E \$48 Under the influence of a linear salt gradient, a series of peaks of lipolytic activity was obtained, eluting at the same, or similar, ionic strengths as those for the MAC16 tumor extracts.

E ls \$49 The fractions giving the main activity peaks were concentrated and applied to a Sephadex G50 column (Flg 4,B).

E x \$50 Again, peaks indicating active components having molecular weights corresponding to 1.5 kd and 0.7 kd were obtained.

E lr \$51 These **results** suggest that other tumors also produce in detectable quantities the same lipolytic factor produced by the MAC 16 adenocarcinoma, even though there may be no obvious symptoms of cachexia.

a) E-lr b) E-x \$52 Structural elucidation of the cachexia-related lipolytic factor awaits its further purification, but preliminary studies have shown it to be stable to heat (90¡C for 15 minutes), acid (pH <1), RNase, DNase, periodate, trypsin, and chymotrypsin.

E x \$53 The activity is, however, partially reduced by pronase, suggesting that the material may be peptidic in nature and, from its chromatographic characteristics, acidic, thus distinguishing it from the natural lipolytic hormones, which are all basic and are not retained by a DEAE-cellulose column (Fig 2).

a) P at b) P2 \$54 To obtain some information on the possible role of the lipolytic factor in cancer cachexia, we administered partially purified material by intraperitoneal injection to female NMRI mice bearing the MAC13 tumor, which does not induce weight loss (Fig 5).

E \$55 Preliminary experiments showed that administration of the lipolytic factor caused a more sustained weight loss in tumor bearing animals than it did in nontumor-bearing animals.

E x \$56 Although the average food or water intake (+ SE) of animals treated with the tumor lipolytic factor (2.6+0.2g/day and4.8+0.1 mL/ day, respectively) did not differ from that of controls (2.8 + 0.3 g/day and 5.0 + 0.2 mL/day, respectively) treated with material extracted from cultures of L1210 leukemia cells and purified by the same procedure, an average weight loss of 1 g was produced in animals treated with the extract from the MAC16 culture medium.

E lr \$57 This **result** suggests that the tumor lipolytic factor may be responsible for the cachexia.

DISCUSSION

TI \$58 Cancer cachexia is characterized by a marked depletion of host lipid stores, and several workers have suggested that neoplastic cells are capable of elaborating a lipid-mobilizing substance.

E-lr \$59 Evidence for this **activity** was first provided by the decrease in host fat content by the injection of nonviable preparations of Krebs-2 carcinoma into male Swiss mice (17), although the material responsible was not further characterized.

E ve \$60 Using an in vivo assay to determine lipolysis by the release of 14C02 from 14C fatty-acid-labeled adipose tissue, Kitada et al (18) reported that the serum of AKR mice bearing a thymic lymphoma contained a potent lipid-mobilizing factor.

E lr \$61 Such **activity** was also obtained with extracts of the tumor, by the injection of samples of culture medium from the lymphoma cell culture, or from a serum sample of a human patient with

adenocarcinoma (19).

E (here = E-ls?) $62 The factor was reported as a heat-stable protein of molecular weight corresponding to approximately 5 kd, similar to that reported here.

E ve (61) $63 With use of an in vitro assay to measure glycerol release after incubation of rat adipocytes with samples of tumor, lipolytic activity could be detected only after aging of the extracts at lowtemperature for several days (20).

a) E lr b) E $64 Following gel filtration of such **aged, active extracts**, the activity of which was completely destroyed by digestion with trypsin, (E) it was concluded that the lipolytically active substance was formed by aggregation of inactive, small protein molecules.

E lr $65 The **activity** of this material differs from the tumor lipolytic activity described in the present report in both its molecular weight and susceptibility to trypsin.

E lr $66 Another **lipolytic factor**, termed "toxohormone-L," isolated from the ascites fluid of DDK mice with sarcoma 180 or from the ascites fluid of patients with hepatoma, appears to differ from the lipolytic activity reported here, in both molecular weight (65 kd to 75 kd) (21) and in the (E ls) fact that it appears to act indirectly by suppressing food and water intake, thereby promoting anorexia as the main cause of breakdown of adipose tissue and symptoms of cachexia (22).

E x$67 Toxohormone-L, however, was reported to be degraded to low-molecular-weight polypeptides, similar to those reported here, which still retained biological activity.

a) E lr b) E-x $68 It is possible that the low-molecular-weight material described in this **report** is derived from a high-molecularweight protein, although we have no evidence for a high-molecular-weight precursor with biological activity.

E x $69 The cachexia-associated lipolytic factor also differs from TNF-A in a number of respects, particularly in molecular weight and ability to induce lipolysis directly in murine adipocytes (4).

E x $70 Preliminary experiments also suggest a reduction in host body weight without a significant drop in food intake when the material is administered by intraperitoneal injection.

a) E-lr or E ve (70) b) E-x $71 The lipolytic factor described in this **report** is closely related to the cachectic state, since it is not present in the serum of normal subjects, even under conditions of starvation.

E lr $72 We have been able to detect the material only in the serum of cancer patients, and the concentration appears to be proportional to the degree of weight loss (23).

a) E ls b) E-x
$73 There is a marked similarity between the material found in the MAC16 tumor and that found in cancer patients in both charge and molecular weight, (E x) although the actual degree of homology awaits the structural elucidation.

E lr $74 The **ability to adhere to DEAE-cellulose** would suggest that the molecule has a negative charge, a characteristic not displayed by natural lipolytic hormones but similar to the material reported by Kitada et al (19) and to toxohormone-L (21).

E x $75 Moreover, inhibition of lipolytic activity by insulin (12), 3-hydroxybutyrate (13), or the essential fatty acids found in fish oil (24) reduces or abolishes the cachexia.

E lr $76 These **data**, together with the induction of weight loss by intraperitoneal injection of purified material, suggest a major role for the tumor lipolytic activity in the development of cachexia.

E ls $77 The presence of similar lipolytic activity, although at a lower concentration in a related tumor, the MAC13, that does not induce cachexia suggests that the material may be important in maintaining

tumor growth by supplementing lipids, which the tumor is unable to synthesize, with those from host adipose tissue.

E x $78 Thus, cachexia could arise from the overproduction of a material present in all tumors, with this activity explaining why the degree of cachexia bears no simple correlation to tumor burden, tumor cell type, or anatomical site of involvement (25).

E $79 Recent results (Rothwell N, Tisdale MJ Mulligan HD: manuscript in preparation) suggest an increased lipid utilization coupled with an increase in oxygen consumption and an increased activity of brown adipose tissue during the period of weight loss.

E lr $80 This increase in metabolic rate would account for weight loss in the absence of anorexia.

E ve (79) $81 If the lipolytic activity is important for maintenance of structural or regulatory lipids in tumors, then it could represent an important target for more selective antitumor agents directed against solid tumors.

a) E-x b) E-x $81 In this respect, we have noted that materials such as 3-hydroxybutyrate (26) and fish oils (24), which are effective in the inhibition of the cachexia, are also effective in inhibiting the growth of the tumor.

APPENDIX B9_

Posture in cancer research article TPS (author *MT*).

(lexical rephrasings are marked in bold. Sentences are numbered by $)

TITLE
TI $1 Newly identified factors that alter host metabolism in cancer cachexia

ABSTRACT
TI $2 Progressive weight loss is a characteristic feature of malignant diseases and some studies suggest that nearly 90% of patients are affected.

E- x $3 Cachexia is also an important factor in cancer mortality, accounting for 22% of cancer deaths (data from 400 autopsies) as well as being implicated as a contributor to the deaths of many other patients.

a) E- x b) E- x $4 Anorexia is also frequently present in cancer patients but numerous studies have documented abnormalities of host metabolism that indicate that cancer cachexia is not the same as simple starvation.

E- x $5 Moreover, attempts to reverse the wasting phenomenon using total parenteral nutrition, whilst leading to some increases in body weight, have been shown to worsen survival to and decrease tumour response to chemotherapy.

E- i $6 **This** suggests that cancer cachexia could arise from the metabolic effect of the tumour on host tissues, mediated by tumour-produced catabolic factors.

a) E- ls b) E- x $7 Attempts to characterize such **factors** have previously focused on macrophage-derived cytokines, but more recent studies have identified a low molecular weight peptide which is synthesized and released by tumours themselves and whose proposed cachexic functions can be inhibited by a fatty acid.

P - (table) $8 Figure 1 shows the possible interrelationship between tumour and host metabolism that could account for the development of cachexia, whereby catabolic factors produced by the tumour lead to breakdown of host adipose tissue and muscle.

E- lr $9 Protein catabolism may account for **some of the immunological abnormalities** commonly seen in cancer patients.

P al $10 The products of catabolism of host adipose tissue and muscle could be used in two **ways**.

P2 $11 First, they could provide nutrients for the tumour either directly, or indirectly after suitable metabolic conversion in the liver.

P3 $12 Secondly, they could be used by the host to create a new metabolic environment that favours the neoplastic state.

E- x $13 For example, free fatty acids could be metabolized to ketone bodies (acetoacetate and 3-hydroxybutyrate) for use by organs such as the brain, as occurs in starvation.

a) E- lr b) E i $14 Despite massive lipid mobilization the plasma level of these **metabolites** is not elevated in the cachectic state and this may be due to an increased utilization of ketone bodies by the brain, coupled with a decreased glucose consumption (M. J. Tisdale and H. D. Mulligan, unpublished).

a) E-x b) E- ls $15 Since the brain is normally the major consumer of glucose, this would reduce the overall host requirements for glucose, allowing its diversion to the tumour.

E- ls $16 Most tumours have a high glycolytic rate, which could arise either from the altered enzyme pattern, or from the reduced oxygen tension caused by the poor tumour vasculature.

a) E- b) E- $17 Gluconeogenesis from non-carbohydrate precursors has been shown to be elevated in tumour-bearing animals and the increased energy expended in futile cycles such as the Cori cycle, an increased conversion of glucose to lipids and the production of ketone bodies may account for the wasting of body tissues in the cachectic state.

E- x $18 In general, tumours have a poor capacity to synthesize their own lipids.

E- x $19 In addition to providing altemative substrates for host metabolism, lipid mobilization may also be important in supplying fatty acids essential for tumour growth.

E- ls $20 Phospholipids containing linoleic or arachidonic acids in the sn-2 position of the glycerol moiety have been shown to stimulate tumour growths.

E- lr $21 One possible mechanism for this **stimulation** could be inhibition of a guanosine triphosphatase activating protein, which in turn inactivates the ras protein.

E- i $22 If this were indeed the case then inhibitors of the cachectic process would be expected to act as inhibitors of tumour growth.

P ts $23 What are the factors responsible for mediating cancer cachexia?

P2 $24 A factor termed cachectin was isolated from rabbits infected with trypanosomes, which produced weight loss and anorexia in this species.

E- ls $25 Cachectin was shown to be identical in sequence to the cytokine tumour necrosis factor (TNF X) 10.

E- lr $26 Studies on animals injected with either TNF-X or cell lines capable of secreting TNF-X in vivo have shown that this **cytokine** can produce most of the changes associated with cachexia such as weight loss anorexia, muscle catabolism and hyperlipidemia associated with an inhibition of adipose tissue lipoprotein lipase.

E- x $27 However, both animal and human studies have failed to detect elevated levels of TNF-X in the cachectic state, while anti-TNF antibodies delayed but did not prevent tumour associated anorexia.

a) E-lr b) E-x (overrides 2nd E-lr) $27 These **results** suggest that if TNF-X is involved in cachexia, then other **factors** must also be present.

E- lr $28 A second **approach** has been to study tumour-derived factors capable of direct catabolism of host tissues.

a) E- ls b) E-x $29 The experimental murine colon adenocarcinoma (MAC16) is a useful model of cachexia, since it is one of the few experimental tumours that produces weight loss with a small tumour burden and without a concomitant drop in food intake.

a) E- ls b) E- lr $30 Cachexia in animals bearing this tumour is associated with an increased plasma level of a lipid-mobilizing factor (LMF); a similar **relationship** between weight loss and serum lipid-mobilizing activity has been observed in a group of cachectic cancer patients.

E- ls $31 LMF is a small acidic polypeptide that is retained by DEAE cellulose, thus distinguishing it from the basic lipid-mobilizing polypeptides present in normal serum.

E- ls $32 Both anion exchange and exclusion chromatography show that LMF resembles a factor found in the serum of cancer patients as regards charge and molecular weight, and also displays some of the

characteristics of a lipid-mobilizing factor previously reported to be produced by a thymic lymphoma in vitro.

a) E- ls b) E-x (overrides E- lr) $33 If LMF is responsible for the cachexia then not only should the serum level correlate with the extent of weight loss, but inhibitors of this **factor** should also be inhibitors of cachexia in vivo.

P al (overrides E-lr) $34 Two observations support this **hypothesis**.

P2 $35 First, the ketone body 3-hydroxybutyrate is an effective inhibitor of the tumour LMF in vitrol, while in vivo studies both in the MAC16 model20 and in cancer patients with severe (32%) weight loss show a significant increase in host body weight with a ketogenic diet.

P3 $36 Secondly, the polyunsaturated fatty acid eicosapentaenoic acid (EPA), but not other related polyunsaturated acids of the (n-3) and (n-6) series, specifically inhibits lipid mobilization by the MAC16 tumour LMF by preventing cAMP accumulation in response to lipolytic stimuli.

E- ls $37 An isocaloric diet in which carbohydrates were substituted by fish oil reversed cachexia in MAC16 tumour-bearing mice.

E- ls $38 Fish oil contains two important fatty acids: eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA).

E- x $39 Administration of pure EPA to weight-losing tumour bearing animals was highly effective in inhibiting weight loss while DHA was ineffective in this respect.

a) E-x or E- ve (35) b) E- x $40 In addition to reducing the cachexia, both 3-hydroxybutyrate20 and EPA (Ref. 22) also inhibited tumour growth, although the effect on host body weight exceeded the effect on tumour growth.

a) E- ls b) E- x $41 In vitro both EPA and DHA are equally cytotoxic to MAC16 cells, while in vivo only EPA displays antitumour activity, suggesting that the in vivo mechanism is not a direct cytotoxic effect.

E- i $42 This suggests that inhibition of the catabolic activity of the tumour may also lead to tumour regression.

E- ve (35) $43 Once detailed structural information on the LMF becomes available it may be possible to produce a new generation of antitumour drugs aimed not directly at the replicative process, but at tumour products that may be essential for growth in vivo.

APPENDIX B10

Posture in cancer research article CCP4 (author *Y W*).

(lexical rephrasing is marked in bold. Sentences are numbered by $)

TITLE
TI $1 Relationship between the melanin content of a human melanoma cell line and its radiosensitivity and uptake of pimonidazole~

Summary.
TI $2 The intra-cellular uptake of the weakly basic radiosensitiser pimonidazole (PIMO) was determined as a function of the pigmentation of Nal l+ human melanotic melanoma cells in vitro.

E- ls $3 Two experimental conditions were considered: exponentially growing cells (Exp.) and plateau-phase cells (Pl.).

E- ls $4 The melanin content of Nal l+ cells ranged from 500 ~g/g cell weight in exponentially growing cells to 6000 ~g/g in heavily pigmented plateau-phase cells.

a) E- ls b) E- $5 Cells were exposed to PIMO (medium dose, 0.2 mmol/dm3; 58.2 ~g/ml); and the intracellular concentration ranged from 163 llg to 900llg.

E- x $6 However, this increase in the cellular uptake of PIMO was not accompanied by an increase in radiosensitising efficiency.

E- x $7 In comparison, the Ci/Ce for etanidazole (ETA), a radiosensitiser that is uncharged atphysiological pH, remained approximately constant at I for all values of melanin contents.

E- $8 Treatment of Nall+ tumours in vivo with L3HIPIMO resulted in a tumour:blood ratio of about 3 at 30-60 min after administration.

E- x $9 However, at 24 h a grain count of label derived from [3H]-PIMO showed that picnotic areas of tumours contained levels that were some 40 times greater than the background value.

E- lr $10 This **high level of label** was coincident with areas of highest apparent melanin content.

E-x $11 In conclusion, PIMO accumulates in very heavily pigmented melanoma cells present in necrotic zones with picnosis.

INTRODUCTION

TI $12 Most rodent and human xenografted tumours contain hypoxic cells [12, 25, 28], and a few clinical studies have suggested that radiotherapy might be improved by the use of agents such as nitroimidazoles that increase the radiosensitivity of these hypoxic cells [6].

a) E- ls b) E-x $13 The first agents evaluated were metronidazole and misonidazole, but neurotoxicity has limited their use in radiotherapy.

a) E- lr b) E-x $14 A second **generation of hypoxic cell sensitisers** have been developed, among which pimonidazole (PIMO) is particularly interesting since it is preferentially accumulated by tumour cells in vitro 14]

E- x $15 PIMO also accumulates in rodent tumours as indicated by tumour: blood ratios of > I [1 5, 20, 2 1, 24, 33, 37, 44-46] and, similarly, in human tumour xenografts [21, 35] and in human tumours in patients [1, 8, 26], with the uptake being highest in melanomas as compared with other tumours [1, 5, 8,

20, 21, 26].

a) E- x b) E-lr $16 The high tumour-cell concentration of PIMO that can be obtained suggests that the radiosensitising effect of PIMO should also be high, and the results obtained using non-melanotic cells growing exponentially in vitro support this **hypothesis** [30, 40].

E- x $17 In addition, PIMO has generally been found to be effective in non-pigmented rodent tumours in vivo [15, 16, 32, 44].

a) E- x b) E- x $18 However, using human tumour xenografts and clinically relevant drug doses, a radiosensitising effect has been detected in rectal adenocarcinoma HRTI 8 but not in melanoma Nal 1+, although the accumulation of PIMO in the melanoma was particularly high [21, 35].

E- lr $19 We report the results of some studies that may explain the basis for **the lack of effect of PIMO on the Nal l+ melanoma**.

E- $20 Tumours are very heterogeneous, especially with respect to the cell-proliferation kinetics (exponentially growing cells, plateau-phase cells) and to the presence of necrotic zones.

E- lr $21 This **heterogeneity** suggests that the average intra-tumour concentration may not necessarily reflect the concentration of the drug in the clonogenic hypoxic cells, which are the cells essential for tumour radiosensitivity.

E- ve (20) $22 In vitro radiosensitisation and drug uptake have previously been investigated in exponentially growing cells [30, 38 -4 1] .

a) P al b) P2 $23 Therefore, the aim of the present work was, firstly, to measure the uptake of PIMO into exponentially growing and plateau-phase melanotic melanoma cells in vitro and determine whether this might be related to their melanin content and secondly, to determine whether the inter- and intra-cellular distribution of [3H]-PIMO in Nal 1 + tumours might explain the lack of radiosensitisation observed for PIMO in this tumour type.

MATERIALS AND METHODS

COMPOUNDS

TI $24 . Etanidazole [ETA, N-(2-hydroxyethyl)-2-nitroimidazolyl acetamide; Chemistry Branch, Division of Cancer Treatment, National Cancer Institute, USA1 and pimonidazole ~PIMO, o~-[(2-nitro-1 imidazolyl)methyl]-l-piperidine-ethanol hydrochloride; Hotfman-La Roche, Switzerland] were dissolved in minimum essential medium (MEM) supplemented with 20 mM HEPES (pH 7.2) at a concentration of 0.2 mmol/dm3 (42.8 llg/ml for ETA and 58.2 llg/ml for PIMO). ~3H]PIMO [42] dissolved in ethanol.

E- ls $25 The total amount of PIMO injected i. v. into mice via the retro-orbital sinus in a volume of 0.3 ml was 200 mg/g body weight. Control mice were given PBS alone.

Tumour-cell system

TI $26. The melanotic melanoma Nal l+ originated from a human melanoma.

E- lr $27 The characteristics and maintenance of this **cell line** have been described elsewhere [14].

E- $28 Congenitally athymic nude mice were bred and maintained in a defined flora- and pathogen-free colony.

E- ls $29 Details of the mouse breeding and tumour production have been published elsewhere [13, 141.

E- ls $30 Tumours were obtained by injecting 3 x 105 cells into both flanks of mice that had been whole-body-irradiated with 5 Gy 137Cs ~-rays to increase tumour uptake.

E- $31 Plateau-phase cells were obtained 7 days after 5 x 105 cells had been seeded in glass petri dishes (unpublished data).

E- ls $32 A few experiments were performed using exponentially growing cells obtained 2 days after the cells had been seeded.

E- ls $33 For the aerobic study, cells were placed in an incubator (37°C, 5% CO2, 45 min).

E- overlay (33) $34 For hypoxia, open dishes were placed in aluminium chambers and gassed with a humidified mixture of 95% N2 and 5% CO2 (<3 ppm O2) for 45 min.

E- $35 A filtered solution of dithionite-sodium carbonate was placed in the centre of the chambers to remove traces of oxygen [18].

E- ve (24) $36 After undergoing incubation with PIMO or ETA and/or irradiation, the cells were trypsinised and the surviving fraction was assessed by an in vitro colony assay.

E- ls $37 Colonies were fixed and stained with crystal violet (0.25%, w/v, in 80% methanol containing 10% formaldehyde).

E- $38 Cellular uptake and drug concentration in the medium was measured by high-performance liquid chromatography (HPLC) 121, 23].

E- $39 Briefly, the medium was analysed directly and cells were collected with a rubber policeman and stored at -80~C.

E- ls $40 On the day of analysis, the cells were sonicated in water and ETA and PIMO were extracted with acetonitrile water (1 : 1, v/v). The extraction was repeated and the supernatants were pooled and evaporated to dryness.

E- ve (38) $41 The residue was diluted in eluent B containing an appropriate amount of internal standard (Ro 03-1902), and cellular uptake was analysed using a Varian model 5000 chromatograph equipped with a 5-1lm Nucleosil column connected to a Varichrom UV-visible detector at 326 nm.

E- $42 The elution flow rate was 2 ml/min (eluent A, 75% acetonitrile and 25% water; eluent B, 4 mM heptane sulphonic acid, 5 mM dibutylamine and 50 mM glycine adjusted to pH 3); the gradient was: 0 min, 8% eluent A; 10 min, 45% eluent A; 10-11 min, 45%-8% eluent A.

P ts IN VIVO STUDIES.
P2 $43 Tumours were used when they had reached a mean diameter of 9-11 mm.

E- $44 Animals were anaesthetised at various times after drug injection and blood was collected by cardiac puncture into heparinised tubes immediately before tumour excision and frozen in liquid nitrogen.

E- ls $45 Tumours were removed and immediately frozen in liquid nitrogen.

E- x $46 Subsequently, tissue samples were weighed and 100 300 mg was suspended in Optisolve (LKB).

E- ls 47 Tissues were digested by incubation at 55~C for 16 h, 15 ml liquid scintillation fluid (Optiphase, Hisafe-TM; LKB) was added and samples were counted on a Packard Tricard liquid scintillation counter.

E- $48 A set of variably quenched standards were used.

E- i $49 These were prepared by adding a known activity of tritiated water to a set of eight samples containing between 0 and 0.2 ml whole blood digested as described above.

E- $50 Any very highly coloured samples were bleached by incubation at 60° C for 2 h with benzoyl peroxide [0.4 ml, 5% (w/v) in toluene] prior to the addition of scintillant.

P ts $51 HISTOLOGY, AUTORADIOGRAPHY AND MELANIN DETERMINATION.

P2 $52 Tumours were removed at 24 h after [3H]-PIMO administration and were immediately fixed in ethanol: acetic acid (3 : 1, v/v).

E- $53 The tissues were dehydrated and embedded in paraffin and sections (3 llm) were cut and mounted on microscope slides.

E-ls $54 For autoradiography the slides were dipped in llford K2 emulsion and exposed for 2 months.

E- x $55 The slides were then developed in Kodak D 1 9b and stained with hemalum and erythrosin.

E- $56 The amount of radioactivity in the tumour was determined by grain counting using an ocular grid.

E- ls $57 The percentage of cellular areas was identified under a microscope and outlined on photographs of the sections.

E- $58 The areas were weighed and cellular areas were expressed as a percentage of the total areas.

E- $59 Melanin was localized by the method of Fontana-Masson [17].

E- ls $60 For melanin determination, the quantitative colourimetric method of Foster et al. [9, 10] as adapted to normal and cultured retinal pigment cells by Whittaker 1431 was used, with a few modifications.

E- ls $61 Cells were sonicated in water, and non-melanic substances that interfered with the assay were removed by three extractions with 5% trichloroacetic acid, two extractions with cold ether-alcohol (1: 3, v/v) and one extraction with absolute ether.

E- ls $62 The dried residue extracted from 50-100 mg cells was dissolved in I ml I N KOH and then heated to 100° C for 30 min.

E- ls $63 A standard curve was constructed using synthetic melanin (Sigma) dissolved in I N KOH (I - 100 lag/ml).

E- ls $64 Absorbance at 400 nm increased linearly with melanin concentration up to 100 ,ug/ml.

RESULTS

TI $65 A preliminary experiment showed that the melanin content was very homogeneous in exponentially growing cells but varied greatly in plateau-phase cells.

E- ls $66 The melanin content of plateau-phase cells obtained at 7 days after seeding increased by a factor of >6 within 12 h (results not shown).

E- ls $67 The uptake of PIMO into hypoxic plateau-phase Nall+ cells in vitro was compared with that of ETA.

P (table) $68 Figure 1 shows that the intra-cellular concentration of PIMO increased linearly (correlation coefficient, 0.9 P:10 4) as a function of melanin content.

E- ls $69 At the lowest melanin content (500 ~g/g), the Ci/Ce ratio for PIMO was between 3 and 4, which is very similar to that found for exponentially growing Nal l+ cells.

E- x $70 However, in plateau-phase cells in which the melanin content increased, the Ci/Ce value also increased to >17.

E-x $70 In contrast, the intra-cellular concentration of ETA remained fairly constant, with Ci/Ce values being close to I .

E- $71 The radiosensitising effect of 58.2 llg PlMO/ml in hypoxic Nal l+ cells was determined as a function of cellular pigmentation.

P at / P2 $72 Figure 2 shows that there was no significant change in the radiosensitivity with pigmentation in plateau-phase cells following a radiation dose of 10 Gy in the presence of PIMO (Fig. 2a), although the intra-cellular PIMO concentration measured in the more pigmented cells was at least 3-fold that determined in the less pigmented cells (Fig. 1).

E- lr $73 In these **experiments**, no significant change in the plating efficiency was detected as a function of either the pigmentation or the intra-cellular PIMO concentration.

E- lr $74 In Fig. 2 b, these **results** expressed as intra-cellular concentration (Ci) are compared with those obtained for radiosensitisation of exponentially growing cells (melanin content, <500 ~g/g).

E- ls $75 The extra-cellular concentration (Ce) remained constant (58.2 ,ug/ml) in plateau-phase cells, whereas it varied from 29.1 to 174.6 ~g/g in exponentially growing cells.

E-x $76 Clearly, in exponential Nal l+ cells, radiosensitisation by PIMO increases as a function of Ci.

E- x $77 In contrast, no similar increase in radiosensitisation was observed in plateau-phase cells.

E- x $78 Furthermore, the maximal Ci value obtained in exponential cells (437 llg/g) was limited by toxicity, whereas in plateau-phase cells a Ci value of 628 ~g/g was achieved with no loss in plating efficiency.

LIQUID SCINTILLATION STUDIES

P - ts $ 79 [3H]-PIMO radioactivity in the blood and in the tumour is shown in Fig. 3.

E- ls $80 The radioactivity in the blood decreased with time after administration, whereas that in the tumour increased to a plateau. The tumour/blood ratio was relatively constant (2-3) at between 30 and 60 min.

P at HISTOLOGY STUDIES

P $81 The cellular zone represented only 51% of the tumour.

E- ve (64) $82 Figure 4 shows the melanin content, with the highest concentrations of melanin being found in the zones of tumour in which picnotic cells form the largest proportion of cells (Fig. 4b).

E- ve (80) $83 The radioactivity derived from [3H]-PIMO was not evenly distributed throughout the tumour.

P (table) $84 The results of grain counts from two experiments are given in Table l; it is apparent that the necrotic zones containing numerous picnoses were most highly labeled, showing values some 40 times greater than the background level.

E- x $85 In contrast, viable zones of tumour and necrotic regions containing no identifiable cell fragments

or picnotic cells showed substantially lower grain counts.

E- ve (83) $86 Only one-fifth of the total radioactivity was located in cellular zones.

E- ve (82) $88 At the lowest melanin content (500 ,lg/g), the Ci/Ce ratio for PIMO was between 3 and 4, which is very similar to that found for exponentially growing Nal 1+ cells. '

E- x $89 However, in plateau-phase cells in which the melanin content increased, the Ci/Ce value also increased to >17.

E- x $90 In contrast, the intra-cellular concentration of ETA remained fairly constant, with Ci/Ce values being close to 1.

DISCUSSION

a) P ts b) P $91 It has been suggested that weak bases can concentrate in melanin-containing cells [3, 20-22, 31, 35, 36].

E- ls $92 The intracellular concentrations of weak bases such as PIMO [4], RSU 1069 and RSU 1164 [37] are increased at elevated extra-cellular pH, and their much higher uptake into melanotic as compared with non-melanotic cells led to the suggestion that the intra-cellular pH of melanotic melanoma cells is lower than the pH of non-pigmented cells.

E- x $93 However, other mechanisms are likely to be involved in the concentration of weak bases in melanotic melanomas.

a) E- b) E-x $94 The present in vitro results show that the pigmentation is higher in plateau-phase cells relative to exponentially growing Nal 1+ cells; furthermore, when the cells reach the plateau phase, a continuous and rapid increase in pigmentation is observed.

a) E- ls b) E-ls $95 For plateau-phase melanotic cells, the results clearly show that the intra-cellular concentration of PIMO depends strongly on the intra-cellular concentration of melanin: the higher the melanin content, the higher the PIMO concentration (Fig. I).

E- lr $96 PIMO had no greater radiosensitising effect on these **cells** than it had on exponentially growing cells, despite the observation that the intra-cellular concentration of PIMO was higher and the plating efficiency was not modified.

E- lr $97 An association between PIMO and melanin could explain these **results**, especially if the melanin is remote from the DNA [34], thus spatially preventing PIMO from exerting its radiosensitising effect.

E- ls $98 The data for melanoma transplanted into nude mice show that at a short time after its administration, PIMO accumulates in tumours (Fig. 3).

E- ls $99 Histological data derived from tumours excised at 24 h after the administration of [3H]-PIMO show the presence of label to be coincident with areas of high melanin content, which would be consistent with the in vitro results.

E- lr $100 It is unlikely that this **localisation** would be a consequence of passive sequestration onto or association with melanin; rather, it is probably due to selective hypoxia-induced binding in these areas.

E- $101 In nonmelanotic tumours it has been claimed that [14C]-misonidazole [2, 11] and PIMO [19, 29] accumulate in viable hypoxic tumour cells, whereas other investigators have reported that uptake is lower in necrotic tumours 17, 27] or that there is no link between accumulation and necrosis [21, 26].

a) E- ls b) E- x $102 On the basis of our results, it is difficult to determine the influence of the necrotic areas on PIMO accumulation, as these regions were also those in which the melanin content was the

highest.

a) E- ls b) E- lr c) E- x \$103 It must be emphasised that although it was not possible to evaluate the melanin content in the necrotic zones, these **dying cells** probably contained much more melanin than did the plateau-phase cells we studied; indeed, the plateau obtained in the present studies is not a perfect one, since 11% of the cells remained in the S phase (unpublished results).

E- lr \$104 On the basis of the **above-mentioned observations**, it is possible to provide an explanation for our previous results, which indicated an inability of PIMO to sensitise NaI 1+ tumours [35].

E- ls \$105 The accumulation of PIMO in the tumour (liquid scintillation studies) may well have resulted from the accumulation of PIMO in necrotic zones containing picnotic cells of very high melanin content.

E- ls \$106 Its accumulation in these regions cannot influence radiosensitivity; only the PIMO in areas of clonogenic hypoxic cells is important.

E- x \$107 In conclusion, the present results obtained both in vitro and in vivo indicate that PIMO accumulates in very heavily pigmented melanoma cells.

E- lr \$108 The in vivo results also show that the accumulation of label derived from PIMO is higher in the necrotic zones with picnosis than in the cellular areas.

E- lr \$109 The **latter findings** could explain why the intra-cellular accumulation of PIMO is not linked with a radiosensitising effect in melanotic melanoma.

a) P at b) P2 \$110 To explain the lack of efficiency of PIMO in vitro, a mechanism such as the localisation of PIMO far from the DNA close to the melanin content or a very low sub-cellular pH must be supposed.

E- ls \$111 As far as melanotic melanomas are concerned, PIMO is probably not the compound to be used in clinical radiotherapy.

APPENDIX C1

Title salient items from the Wordlist program

NB Some items were mis-scanned in the original corpus. I have marked them [sic].

KEYWORDS RANK	WORD	PSC Titles Freq. in subcorpus	%	PSC Freq. in whole corpus	%	Chi sq.	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	SUBSTITUTED	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	
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	

APPENDIX C1 (Cont.)

Title salient grammatical items from the Wordlist program

RANK	WORD	PSC Titles Freq. in subcorpus	%	PSC Freq. in whole 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

APPENDIX C2

Abstract salient items from the Wordlist program

RANK	WORD	PSC Abstracts Freq. in subcorpus	%	PSC Freq. in whole corpus	%	Chi sq.	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

Abstract salient grammatical items from the Wordlist program

RANK	WORD	PSC Abstracts Freq. in subcorpus	%	PSC Freq. in whole 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

APPENDIX C3

Introduction salient items from the Wordlist program

RANK	WORD	PSC Intros. Freq. in subcorpus	%	Freq. in whole corpus	PSC %	Chi sq.	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	TRANSPOSONS	12		16		27.3	
51	INCLUDE	21		47		27.2	0.000

APPENDIX C3 (Cont.)

Introduction salient grammatical items from the Wordlist program

RANK	WORD	PSC Intros. 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

APPENDIX C4

Methods salient 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
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

Methods salient grammatical items from the Wordlist program

RANK	WORD	PSCMethods 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

APPENDIX C5

Results salient items from the Wordlist program

RANK	WORD	PSCResults Freq. in subcorpus	%	PSC Freq. in whole corpus	%	Chi sq.	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	GST	48		71		32.3	0.000
51	POSITIVE	124	(0.1%)	282		31.9	0.000

APPENDIX C5 (Cont.)

Results salient grammatical items from the Wordlist program

RANK	WORD	PSCResults 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

APPENDIX C6

Discussion salient items from the Wordlist program

RANK	WORD	PSCDiscussion 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
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	RECEPTORS	51		85		30.4	0.000
51	DUE	108		252		29.5	0.000

APPENDIX C6 (Cont.)

Discussion salient grammatical items from the Wordlist program

RANK	WORD	PSCDiscussion 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