

On the Phraseology of Grammatical Items in Lexico-grammatical Patterns and Science Writing

Christopher Gledhill

▶ **To cite this version:**

Christopher Gledhill. On the Phraseology of Grammatical Items in Lexico-grammatical Patterns and Science Writing. Paul Thompson; Giuliana Diani. English for Academic Purposes: Approaches and Implications, Cambridge Scholars Publishing, pp.11-42, 2015, 978-1-4438-7439-7. hal-01220012

HAL Id: hal-01220012

https://u-paris.hal.science/hal-01220012

Submitted on 28 Jun 2021

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

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

CHAPTER ONE

ON THE PHRASEOLOGY OF GRAMMATICAL ITEMS IN LEXICO-GRAMMATICAL PATTERNS AND SCIENCE WRITING

CHRISTOPHER GLEDHILL
UNIVERSITÉ PARIS DIDEROT, FRANCE

1. Introduction

In this chapter I examine the role of grammatical items in lexico-grammatical patterns (or 'LG patterns', for short). In previous work I examined the collocational patterns of individual grammatical items in a particular genre (the cancer research article, Gledhill 1995, 2000a, 2000b). In these studies, I demonstrated that individual functional words (such as 'and' in Titles, 'but' in Abstracts, 'to' in Introductions and so on) have a non-random distribution in these texts, since these words are 'statistically salient' or 'key' in these different parts of the research article. I then went on to examine in detail the phraseological behaviour of these items in each subsection, arguing that, contrary to what one might think, each grammatical item enters into a very restricted, predictable set of phraseological patterns, according to the type of text being analysed. It is often thought that grammatical items do not enter into collocational relations, since they can 'be used anywhere' and thus can 'collocate with anything'. However, one of the findings of my work has been to argue that grammatical items have a highly restricted phraseology in specialised discourse, a feature which makes them ideal targets for analysis, since an analysis of the distribution and behaviour of function words can in effect be seen as a preliminary analysis of the fundamental stylistic and phraseological properties of a particular text type.

In this chapter, I use similar methods and I make similar claims. However, my focus here is somewhat different. In this study, I concentrate on longer stretches of wording, and in particular I am interested here in examining discontinuous stretches of text or what Renouf and Sinclair (1991) have called "collocational frameworks". A discontinuous stretch of text is a short sequence of words such as *a(n) * *-ed in* in which only a few selected kinds of lexical item (marked by *) can fit meaningfully into the lexical pattern (in this case '*A substance found in... A cytokine implicated in'...* etc.). What I find interesting about such sequences is that they often correspond to short extracts of very highly specialised discourse. The main hypothesis I wish to test here is that by searching for a given sequence of grammatical signs, it is possible to identify a regular pattern of discourse, provided that this pattern is looked for in a relatively coherent body of texts (i.e. an electronic archive or corpus). Furthermore, I would suggest that by looking at discontinuous sequences in this way, the corpus analyst should be able to identify some of the most characteristic features and functions of that genre in an efficient and systematic manner. In this chapter, I look at examples taken from a corpus of research articles and their corresponding abstracts (referred to below generally as RAs), as well as 'journalistic accounts' of the same research (referred to as JA). However, as I point out in the data analysis below, although many discontinuous stretches are highly regular and recurrent in the particular texts I am interested in here, some kinds of writing (notably scientific journalism) also involve 'hybrid' patterns, i.e. an original blending of two or more patterns that are commonly found in other types of discourse.

In this contribution I use the term "grammatical item" (also known as closed-class item, function word, small word, stop-word, etc.) in contrast to "lexical item", and I assume that the difference between the two is that grammatical items belong to a relatively closed class of high-frequency, polyvalent words with relatively abstract meanings, such as *auxiliaries, conjunctions, determiners, grammatical adverbs, prepositions* and *pronouns*. Until recently, grammatical items as an entire class have not received much attention in English for Specific Purposes and corpus-based genre analysis. Indeed, it has often been supposed that grammatical items are of little interest in text linguistics because they can "go with anything" (i.e. collocate with any lexical item, appear in a vast range of grammatical structures, occur in a uniform way across almost all text types, etc.). Thus in the early days of lexicometrics and Natural Language Processing (NLP), specific procedures were developed to filter these words out from the analysis (for example, Smadja 1993 introduced a well-known method of extracting collocations from the Hansard bilingual corpus, but only after

a process of automatically filtering out the function words). To a certain extent the concept of the stop-word is still widespread, and entirely understandable: analysts are interested in getting at what they consider to be important data (equating this with lexical items, content-words, terminology), while the large quantities of apparently ubiquitous grammatical items which fill most types of text appear to be redundant 'noise'.

However, since the advent of corpus linguistics (and particularly the use of corpora by grammarians), there has been a growing body of evidence to suggest that grammatical items enter into collocational relations that are just as interesting and revealing as lexical items, and thus have an important role to play in the phraseological patterning of texts. Renouf and Sinclair (1991) and Renouf (1992) most notably argued that grammatical items are the building blocks of idiomatic language, and coined the term "collocational framework" to refer to such productive sequences as *a(n) X of (a [dash, handful, smattering] of)*. More recently, and in a way that mirrors the current move to rehabilitate 'junk DNA',[1] some researchers in NLP now recognise the importance of functional words in automatic text recognition, terminology extraction and other corpus-based applications (Meyer 1988; Riloff 1995; Vergne 2004). Similarly, collocational frameworks and related notions such as "bundles", "clusters", "n-grams" and so on, have become an accepted part of the descriptive apparatus of corpus-based applied linguistics, and many researchers have examined the distribution and collocational behaviour of specific grammatical items as they occur in specialised corpora (Luzón Marco 1999, 2000; van der Wouden 2001, 2007; Biber et al. 2004; Cheng et al. 2008; Hyland 2008; Bordet 2011) as well as the role of grammatical sequences and frameworks in discourse analysis, language learning and evaluation (Hasselgren 2002; Groom 2005, 2010; Scott and Tribble 2006; Biber and Barbieri 2007; Lee et al. 2008).

But although the study of function words has now become an accepted part of the corpus-based approach, the notion that grammatical items can be the focus of phraseological analysis still requires some theorisation within a broader analytical framework. In addition, it seems to me that for most observers, the idea that grammatical items can be the starting point for textual analysis is still not obvious. Therefore, before embarking on an analysis of discontinuous lexico-grammatical patterns, I set out in the first half of this chapter some of the arguments for studying grammatical items from the point of view of corpus-based genre analysis.

2. Why study grammatical items?

2.1. Grammatical items have collocations

Not all analysts accept that grammatical items enter into collocational relations. For example, in one well-known British dictionary of linguistics, "collocation" is defined as the co-occurrence of lexical items, while grammatical items are explicitly stated as having no collocational relations:

> collocation, n. A term used in lexicology by some (especially Firthianlinguists) to refer to the habitual co-occurrence of individual lexical items […] Some words have no specific collocationalrestrictions - grammatical words such as *the, of, after, in* [...] Another important feature of collocations is that they are formal (not semantic) statements of co-occurrence [...] (Crystal 2008, 86-87)

This definition seems to represent a commonly-held view among linguists. However, I feel that Crystal rather misrepresents the way Firth (1957) would have understood the term, or at least as Firth's successors understand it. From a Firthian point of view, every single sign in the language (whether lexeme or morpheme) has a consistent and contrastive context of use. In other words, each sign is used in a consistent and contrastive set of linguistic co-texts (e.g. the preposition *of* typically has as a complement the noun *course*) and each sign is used in a consistent and contrastive set of situational contexts (e.g. *of course* tends to be used as an informal, concessive adjunct). The term "context" is clearly central to this approach, and in typical Firthian fashion it is used to indicate three things at the same time: a) the "the co-occurrence of forms within the same stretch of text" (usually what is meant by "co-text"), b) the immediate "context of situation", and c) the broader "context of culture". The point about co-text and context is that they essentially shape the meaning of the linguistic sign, since signs are mutually dependent on their typical collocational partners in discourse, as Firth puts it:

> The collocation of a word or a 'piece' is not to be regarded as mere juxtaposition, it is an order of mutual expectancy. The words are mutually expectant and mutually prehended. (Firth 1957[1951], 181)

This kind of definition sets itself against an "essentialist" or "semantic trait" approach to meaning. Thus, it is claimed that the meaning of a word such as *of* can only be seen as a composite of its particular uses, which

depend partly on its co-occurrence with *course* in *of course* and partly on other uses, such as its co-occurrence with a nominal post-modifier referring to people in terms of subjective, usually high-mindedly positive qualities (*a man / woman of [action, honour, humility*, *steel, quality]*) and so on. Sinclair (1991) argued that these and the other typical lexico-grammatical patterns associated with *of* contribute to our general understanding of this rather idiosyncratic preposition. As he puts it:

> Most everyday words do not have an independent meaning, or meanings, but are components of a rich repertoire of multi-word patterns that make up text (Sinclair 1991, 108).

2.2. Grammatical items have different distributions in different text types

Many linguists would agree that different varieties of language exploit different lexical and grammatical resources, an assumption which lies behind the multi-factorial method of register analysis developed by Biber et al. (2002). Yet it is surprising to see how few studies of a particular text type begin by setting out the relative distribution of grammatical forms, especially grammatical items. In this chapter, I examine the role of grammatical items in their local contexts (as the most recurrent, pivotal items in lexico-grammatical patterns). But before examining their local role, it is important to realise that grammatical items (indeed all items, whether functional or lexical) have a particular distribution in different texts, and even within different subsections of texts. The reason for this is that that, as shall be shown in the following sections, if the occurrence of a particular grammatical item is statistically more 'salient' or 'key' in a particular text type or subsection of a text, this is because this item is pivotal in those lexical patterns which have an important role or discourse function to play in that text. One example of this will suffice here: in Gledhill (2000) I showed that the word 'to' is an statistically significant item in cancer research article introductions. One reason for this, it would seem, is that 'to' is a pivotal element in post-posed attributive clauses such as *'it is (important, necessary, possible) to (assess the cell differentiation at this stage, construct a series of structures, identify TAAs, repeat measurements)...*' but also in passive non-finite projecting clauses such as '*(HPV 16 E6, hyperphasis, metabolic inc-cells, is (known, likely, thought...) to (be involved, be a major factor, determine celle cycle) in' etc*.' (from the Pharmaceutical Sciences Corpus, Gledhill 2000, 151). Examples such as these demonstrate three principles: 1) grammatical items collocate with other lexical and grammatical items (note the discontinuous

sequence *it is X to* in the first pattern and *X is Y-ed to Z in* in the second pattern), 2) each lexico-grammatical pattern expresses a specific discourse function (in the first case, 'strongly stating the case for a clinical methodology' and in the second case 'tentatively proposing a biochemical explanation'), and 3) the most systematic way of identifying these patterns is, in my view, to compare the relative distribution of grammatical items across different text corpora.

As mentioned in the introduction, in Gledhill (1995, 2000a, 2000b) I attempted to show that grammatical items have a particular distribution across the different rhetorical sections (Titles, Abstracts, Introductions, etc.) of 150 research articles (RAs) in the field of cancer research (the Pharmaceutical Sciences Corpus, PSC). At the time, I did not have access to a specialised reference corpus in English but later I used the British National Corpus (BNC) as a reference corpus. For example, the following tables (1, 2 and 3) present the results of a "keyword" comparison using AntConc (Antony 2002). In order to obtain these results, AntConc first creates a word list for each corpus (the BNC has 100,520,565 tokens with 448,005 types, and the PSC has 1896869 tokens with 48,537 types).[2] AntConc then calculates a score for each word by comparing a particular item's percentage chance of occurring in the study corpus as opposed to its chances of occurring in the reference corpus. For example, in the PSC *were* occurs 17,968 times divided by 1,896,869 tokens (=0,0961 or roughly 0.96%), whereas in the BNC *were* occurs 306,801 times divided by 100,520,565 tokens (=0,00305 or roughly 0.30%). Such a large percentage difference shows the extent to which certain function words, such as *were,* can have consistently different distributions across text types. It is of course important to be able to judge what is meant by "large percentage difference", and the "Keywords" module of AntConc is an important stage in this analysis. Keywords compares the relative percentages of items as they occur in the study corpus and reference corpus, and then assigns a rankto each item according to a statistical test (for details see Scott and Tribble 2006).[3]

The tables set out the first 20 keywords in the PSC (ranked by decreasing keyword score) as compared with the BNC (Table 1) and the first 20 key grammatical items in the PSC as compared with the BNC (Table 2).

Rank	Item	Freq. in PSC	Keyword score(vs. BNC)	Rank	Item	Freq. in PSC	Keyword score (vs. BNC)
1	&	6071	48432.645	11	Table	2231	8976.632
2	patients	8563	36825.897	12	clinical	1812	8445.988
3	et	4540	22956.671	13	cell	2165	8335.861
4	al	4544	21851.859	14	min	1447	7897.745
5	study	5791	19067.584	15	Fig	2126	7643.490
6	cells	3911	16769.641	16	cases	3053	7609.082
7	were	17968	15873.864	17	patient	2249	7373.035
8	results	3664	11613.943	18	studies	2467	7307.815
9	treatment	3212	10486.346	19	significant	2410	6703.943
10	of	82182	9404.549	20	tissue	1373	6402.795

Table 1. First twenty keywords in the Pharmaceutical Sciences Corpus (PSC) vs. the BNC.

Rank	Item	Freq. in PSC	Keyword score(vs. BNC)	Rank	Item	Freq. in PSC	Keyword score (vs. BNC)
7	were	17968	15874.251	228	these	3859	1652.307
10	of	82182	9404.238	236	However	1702	1627.852
38	with	21063	5318.315	247	due	1203	1595.017
45	in	47809	4915.890	260	may	4165	1529.965
58	and	61723	3824.316	261	The	16217	1528.736
91	during	2581	2971.304	347	Therefore	461	1250.765
117	between	4148	2521.590	354	or	9940	1230.950
124	In	6005	2464.055	361	after	3479	1213.382
163	vs	396	2114.845	364	was	20876	1208.869
189	versus	456	1888.239	519	both	2313	921.387

Table 2. First twenty grammatical keywords in the Pharmaceutical Sciences Corpus (PSC) vs. the BNC.

Analysts familiar with keyword lists will have little difficulty in interpreting these data. The keywords in Table 1 show some of the major textual features of the PSC (such as the presentation of *data* in *Tables* and *Figures*) as well as the topical preoccupations of the PSC (the nominal expression of material processes, e.g. the *study* of *cells, cases* or *groups* of different *ages* and at different *times*, the *treatment* of *patients*, the *use* of drugs / *treatments* and the verbal expression of communicative or perceptive processes *significant results found* or *reported*). Similar

comments can be made about the key grammatical items in the PSC (Table 2): they are predominantly prepositions and coordinators (*of, and, or,* typically involved in elaborate nominal post-modifiers in the PSC) or prepositions involved in adjuncts / post-modifiers expressing cause (*due to*), accompaniment or manner (*with*), temporal extent (*after, between, during, in*) and comparison (*between, versus, vs.*). Table 2 also shows the typical markers of cohesion which we might expect to find in elaborate written discourse, such as pronouns / determiners (*both, The, these*) and sentence-initial conjuncts (*However*, *Therefore*). Other items are perhaps less obviously typical of written discourse, but Table 2 shows that they are salient in science writing: *may* (the preferred modal verb for "hedging", especially in Discussion sections of the PSC) and *were* (usually an auxiliary expressing the past passive in Methods sections).

As mentioned above, my initial description of the PSC was an intra-varietal analysis, conducted in order to establish the main differences between the different rhetorical sections of the research article and the PSC corpus as a whole (Titles, Abstracts, Introductions, Methods, Results, Discussions). I shall not repeat these data here, but for illustrative purposes, the following Table 3 sets out the main results for the first ten key grammatical items across each sub-section of the PSC:

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

Table 3. First ten grammatical keywords in the six main sub-sections of the PSC.

Although general comparisons (Tables 1 and 2) give a good idea of the general features of the PSC, Table 3 shows the extent to which there is also much variation within the research article genre itself. In fact there is so much internal variation that some items which are statistically salient in their respective sub-sections of the PSC, are also more typical of the

general language (BNC) when compared with the PSC as a whole (in particular: the pronouns *we, our* in Introductions and Discussions, the modal *can* which is only salient in Introductions, the item *to* which is also salient in Introductions, as mentioned above). In Table 3, I have indicated the items which stand out in relation to the other sub-sections of the RA (by underlining). I shall not go into a detailed analysis of these data here. It is sufficient to note that over half the keywords in the Introductions (*been, has, such, can, it, to*) and Methods (*were, at, then, each, from, with*) are only specifically "key" in these sections, a result which suggests that these sections have a specific phraseology which is quite unlike the rest of the research article (although these items are of course not exclusive to these sections).

2.3. Grammatical items are pivotal elements in lexico-grammatical patterns

In the previous section, I showed that grammatical items do not have an even distribution across text types, and that their distribution varies considerably, even within the same text type. In this section I argue that the identification of "key" grammatical items can be seen as a useful first stage in the search for longer stretches of phraseology. Some authors (notably Hunston and Francis 2000) use the term "lexical pattern" to refer to regular multi-word units which do not necessarily correspond to the traditional constituents of the clause. In this chapter, (and elsewhere, Gledhill 2011) I refer to such sequences as "lexico-grammatical" (LG) patterns, in an attempt to make it clear that in any multi-word phrase at least one grammatical item (or grammatical structure) is a permanent or "pivotal" element around which the rest of the phrase is built.

In order to illustrate this notion, let us return to the particular case of Abstracts in cancer research articles (in thePSC there are 400 Abstracts = 123,296 words or 6.5% of the corpus). As mentioned above, the first ten grammatical keywords in this sub-section are (in order of rank) *but, these, of, there, in, was, that, did, who, both.* Out of context, it is not clear what patterns of usage these items might represent. An item such as *that* can be used in many different lexico-grammatical contexts (conjunction, pronoun, determiner etc.), and it is therefore necessary to analyse each item separately, not only within Abstracts, but also contrastively, in the rest of the research article. This is not an easy task, not least because the analysis of grammatical items usually generates a vast amount of data. However, I would suggest that the task is simpler when looking at a specialised genre than for the general language. For example, in the 1st edition of the Collins

Cobuild dictionary (Sinclair 1995), there are 19 entries for *of* (not including idiomatic uses), whereas in the PSC (Gledhill 2000,142-149) the number of patterns varies between 3 (Titles, Abstracts) and 5 (Introductions). Even so, it is still difficult to represent this kind of data, and I will not repeat the analysis of each of these patterns here, largely because this type of detailed analysis requires long lists of concordances. However, in Gledhill (1995) I suggested one way of resolving the problem of data representation, which I called the "collocational cascade". An example of this is set out in the following figure:

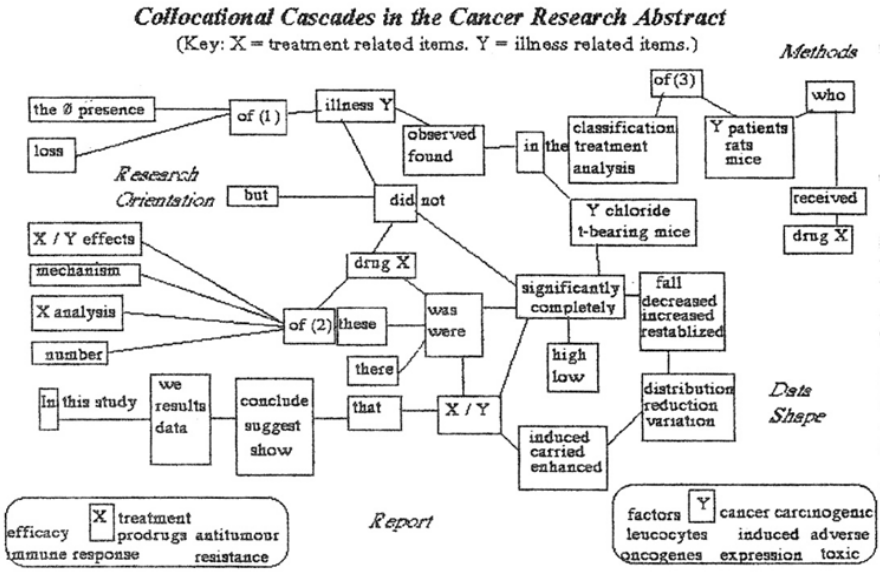


Figure 1. Collocational cascades in Cancer Research Abstracts.[4]

I would argue that "collocational cascades" are an efficient way of summarising the most salient phraseology of a particular text type. Thus in Figure 1 we can see that the outstanding or salient phraseology of Abstracts involves a specification of the general shape of data (as evidenced by lexico-grammatical patterns involving *in, of, that, there*), statements about who or what was affected by various treatments (the LG patterns around *in*, *of, who*) and the extent to which an effect was or was not observed (the LG patterns involving *but, did, not, was, were*). Unlike diagrams representing collocational networks (Williams 1998),

collocational cascades do not represent a formal or statistical relationship between lexical items. Rather, the cascade is broadly meant to be read from left to right, as an informal representation of interlocking lexico-grammatical patterns which, as the cascade metaphor suggests, fall or lead on from one choice of expression into another further on in the clause. In Figure 1, each grammatical item in the cascade (*but, did* + *not, in, of, that, there, these, who,* but not *both*) is linked to one or more of the main patterns as observed in the Abstracts sub-corpus (not counting of course the many sub-patterns or variants of these patterns). We can also see in the diagram that some items appear more than once. Thus, the diagram shows that there are 3 (main) patterns involving the preposition *of* in Abstracts: (1) a quantification (*loss / presence*) of a (usually post-modified) disease-related item (*cancer cells, carcinogenic factors, leucocytes...*), (2) an observation, quantity or facet (*amount*, *analysis*, *effect*) of a treatment-related item *(antitumour response, immune response, prodrug*), and (3) an extended pattern involving a reduced relative clause (expressing a mental / empirical-oriented process: *found*, *observed*) qualified by a complex nominal (expressing a research-oriented process: *in the* + *analysis / classification / treatment of* + *disease-related item*). While these patterns are clearly prevalent in Abstracts, they are also clearly typical of the complex (post-modifying) nominals analysts have come to expect in academic and scientific writing in general. In addition, in the following section, we see that pattern 4 has a slightly different realisation in journalism (examples 4g, 4h and 4i).

One of the defining features of collocational cascades is that the patterns they represent are all related (each grammatical and lexical item is linked indirectly to one or more other items, creating a complex, although sometimes incomplete chain of patterns). I would claim that the grammatical items in the cascade are "pivotal", that is to say they are used consistently in each of these patterns. The lexical items on the other hand represent a "paradigm", in thatthey usually represent semantic abstractions or families of related lexical items rather than specific examples (as in *of* pattern 1: *loss / presence of* + Y, where Y is the name of a specific disease-related item such as *leucocytes*). In addition, collocational cascades have a certain directionality. What I mean by this is that the cascade as a whole represents the general way in which information is structured within Abstracts. For example, expressions and phrases which are research- or report-oriented (*In this study, we conclude that...*) as well as empirical observations (*loss / presence of item Y*) appear in theme (clause-initial) position in this diagram, whereas clinical methods (*item Y*

who received item X) and results (*did not significantly fall / increase...*) appear in rheme / clause-final position.

I make no claim here about the linguistic status of collocational cascades; I primarily see them as a way of depicting the outstanding phraseology of a particular text type. However, I would suggest that this kind of representation does capture something of the social or psychological reality of this kind of formulaic language, in which members of the discourse community recognise that they write in "chunks" and "formulae", and claim to "skim" research articles before deciding whether to pore through them line-by-line. These notions, as well as non-linear processing, predictive text analysis and lexical priming, have become important themes in applied linguistics (de Cock 1998; Simpson 2004; Hoey 2005). However, I will not dwell on these issues here. The more general point I am making is that lexico-grammatical patterns are significantly recognisable within a particular text type, and that LG patterns can be seen as parts of a broader set of interlocking collocational cascades within that particular type of discourse.

3. Extended lexico-grammatical patterns

So far, I have argued the case for seeing grammatical items as key elements in the corpus-based analysis of genres, since these words are the pivotal building blocks of lexico-grammatical patterns. In this section, I attempt to establish whether longer stretches of LG patterns can be identified on the basis of corpus analysis, and my particular focus here is on patterns involving extended (and usually discontinuous) sequences of grammatical items. Previous work on collocational frameworks has usually focused on the collocation of two grammatical items within a pre-defined window of words (Renouf and Sinclair 1991; Cheng et al. 2008; Groom 2010). Here, on the other hand, I am interested in identifying patterns in which at least one grammatical sign is bound between two other grammatical items, such as *the* * * *of* **s* (where * is a wild-card representing one lexical item of any length, * * are contiguous lexical items, and [* *of*] or [**s*] stand for lexical items with an intervening grammatical item or an attached grammatical morpheme). My hypothesis is that discontinuous sequences of grammatical items are particular to specific genres, and that when they can be observed with sufficient regularity, they provide good evidence for the existence of the typical lexico-grammatical patterns of that text type. In other words, given two sequences, such as a) and b) below (both sequences are complete sentences), it should be possible to predict whether they belong more or

less to the typical discourse of a research article (RA) or a journalistic article (JA), and within these sequences it should be possible to identify those patterns that are typical of the genre and those which are 'merely' local innovations:

a) * * * is a * *of * *s and some *s * that *ly * of *s are *ed.
b) *s * a * *er to *ing an * * * and * * of * of the most * and * *s.

In order to test the 'extended pattern' hypothesis, I have examined a sample of complete sentences taken from research articles on cancer cachexia (two of which authored or co-authored by Professor Michael Tisdale, Aston University, and both included in the Pharmaceutical Sciences Corpus, PSC) and from a selection of journalistic articles which all refer to this research as a "breakthrough".[5] In the following sections I look at the initial sentences from two research articles written by Michael Tisdale and his colleagues (RA1: *Trends in Pharmaceutical Sciences* and RA2: *Journal of the National Cancer Institute*) and I then look at the initial sentences from journalistic accounts which refer to this research as a 'breakthrough' (JA1: *The Daily Telegraph,* JA2: *The Independent,* JA3: *The Guardian* and JA4: *The Birmingham Post*). In each case, I attempt to find the sequence in two corpora: the British National Corpus (BNC) for the general language and the Pharmaceutical Sciences Corpus (PSC) for scientific discourse. For each sequence, it is possible to multi-word searches of the form *the * of ** (although AntConc often interprets the * symbol as more than one word). When a sequence turns out to be impossible to find (which is the case for most examples above approximately 10 signs), I then search for increasingly shorter extracts.

(RA1) *Trends in Pharmaceutical Sciences*[6]

The first sentence of this research article reads as follows:

(1)
Progressive weight loss is a characteristic feature of malignant diseases, and some studies suggest that nearly 90% of patients are affected. (RA1)

Here is the sequence of grammatical items used in the search:

* * * is a * *of * *s, and some *s * that *ly * of *s are *ed. (RA1)

I have included the sign *is* in the search sequence, even though it is used here as a lexical, copula verb (the verb *are* in the second half of the extract is an auxiliary, a more *bona fide* grammatical item). One

justification for doing this is that if *is* is not included, a search for the sequence * * * * *a* * **of* produces too many hits and includes many irrelevant patterns. Using *is* to narrow down the search, I find 264 examples of the sequence: *is a * * of *, and* in the BNC. Many of these examples do not include a clause break before *and,* as in *this show is a triumphant affirmation of life and vitality*, and very few involve a complex nominal (as we have at the beginning of extract 1). Only 11 BNC examples are structurally close (but still not exactly matching) extract 1. However, it is interesting to note how similar these examples are topically to extract 1, all involving highly technical subject nouns and an attributive clause which either defines or evaluates the subject as a more general "cause", "source", "method", "product" etc.:

(1a)
COPD **is a leading cause of** morbidity and mortality worldwide, and results in an economic and social burden that is both substantial and increasing. (BNC)

(1b)
Pamidronate, a second-generation bisphosphonate, **is a potent inhibitor of** resorption, and has been successful in the treatment of TIH. (BNC)

(1c)
This dividing technique **is a useful method of** increase, and works well, provided each piece has some root and some dormant buds or young shoots. (BNC)

A similar search in the PSC of course finds RA, plus 10 other examples, this time with complex nominal structures in subject position. In Gledhill (2000) I found that the sequence *is a* is a salient sequence in research article introductions. It seems that this usage is simply one realisation of a more general pattern in academic writing, in which *to be* introduces an attributive complement in the present tense and has the discourse function of expressing explicit evaluation. As can be seen in the following PSC examples, the phraseology of these patterns is similar to that of the BNC, except that the complement typically refers either to a key biochemical agent / participant, or to a source / cause from the point of view of the observer (example 1f):

(1d)
The present inhibition studies show that MAMC **is a competitive inhibitor of** dextromethorphan, and vice versa. (PSC)

(1e)
The oncogenic Bcr-Abl tyrosine kinase **is a potent inhibitor of** apoptosis, and it is retained exclusively in the cytoplasm of transformed cells. (PSC)

(1f)
This reliance on symptomatic presentation and recall of poorly defined symptoms **is a significant source of** bias and results in underestimates of the true incidence of each event. (PSC)

The second half of extract 1 involves a reporting clause (a projecting clause, in systemic functional terms: Halliday and Matthiessen 2004). Although it isnot possible to find the precise sequence *and some *s * that *ly * of *s are *ed* in the BNC, it is possible to find the first part of the structure *and some *s * that* (27 examples). Most of these correspond to a projecting clause with a fairly predictable set of subjects (*analysts, accounts, commentators, estimates, reports*) and verbs (*argue, believe, claim, indicate, suggest*). The PSC does not contain any clause of this type introduced by *and*, but does have 20 examples of projecting clauses of the form *some *s * that* (to cite just one example *While some authors recognize that acute post-operative airway obstruction is common...*). The second half of this sequence (the projected clause *that *ly * of *s are *ed*) is more problematic: the only matching sequence which can be found in the BNC is **ly * of *s are *ed*, and none of these examples have the same grammatical structure (except for some marginally related examples, such as *eventually sets of compounds are perceived... only representatives of parties are elected*). A search of the PSC however produces 82 hits of the form *modal Adv A N of N*, although few of these occur as a projected clause. It is also notable that while the sequence **ly * *s of* matches almost exactly the statistical analysis of clinical methods or the reporting of results that we have in extract 1, the verb forms used are more likely to be past active or past passive:

(1g)
Approximately one third of respondents were within 10% of the Australian rate (PSC)

(1h)
it could be expected that **significantly lower doses of antioxidants could be used** in future prevention studies (PSC)

(1i)
Hence, **significantly greater levels of eosinophils were recovered** from the lungs of antigen-challenged mice following prior treatment (PSC)

Thus the only major difference between these examples and extract 1, is that our original sentence uses the present tense, which is more usual in Introductions. It is tempting to suggest that the originality of extract 1 (in contrast to the PSC, and perhaps the general discourse of science writing) is that it adopts a canonical introductory style initially, but then shifts into the phraseology of results reporting. This sort of shift must surely be related to what follows in the argumentation structure of this article. Of course this comment highlights the limits of analysing sentences in complete isolation from the rest if their original context.

(RA2) *Journal of the National Cancer Institute*[7]

The first sentence in this research article reads as follows:

(2)
Recently, considerable attention has been directed toward the isolation and identification of the factors responsible for the complex metabolic changes associated with cancer cachexia. (RA2)

Here is the same sequence as a discontinuous framework:

*ly, * * has been *ed toward the * and * of the *s * for the ***s* ed with * * (RA2)

The first few signs in this sequence **ly, * ** provide too many hits in the BNC. But a direct search for *has been *ed toward* does not give any results. However, when I change the verb form (*are / was / were*) and the preposition (*towards,* which is more typical of British English), I find a large number of structurally similar examples (approx. 250). The verbs occurring in this pattern have a consistent meaning (*aimed, directed, orientated*), and the subject nouns refer consistently to research-oriented processes or general cognitive processes (*activity, effort, study*):

(2a)
At Sunbury, XTP's **activities are directed towards** helping the business add value (BNC)

(2b)
most of these **efforts were directed towards** reducing non-oil imports, which had damaging effects on domestic production. (BNC)

(2c)
Early evaluation **studies were directed towards** the use of specific media (BNC)

The same search of the PSC reveals 18 examples of *(be) *ed toward* and 14 examples with the spelling*(be) *ed towards.* As with the BNC data, no examples in the present perfect can be found. The general pattern relates to research activity (around half of the time involving the verb *directed*, the same pattern as that of extract 2 (examples 2a-c). A variation on this pattern (examples 2d-f) involves more observation-oriented verbs (*shifted, skewed, weighted*), relating to the changing 'shapes' of empirical data:

(2d)
Future **research** in this area should be **directed toward** tracking the intracellular signal transduction pathways (PSC)

(2e)
Particular **interest** was **directed towards** a careful topographical analysis of the obtained proliferation data within different sites of the vessel wall. (PSC)

(2f)
The maximal luciferase activity was the same for the three steroids, but the **curve** obtained with testosterone was **shifted toward** a higher concentration of ligand. (PSC)

The second regular pattern to be observed in extract 2 involves the sequence: *the *s * for the ***s*. The BNC has 56 examples of this sequence, although many of these are post-modified nominal groups involving a past participle (of the type: *tested for*). A small handful of examples resemble extract 2 more closely, with a post-modifying epithet such as *accountable, available, responsible:*

(2g)
the persons **accountable for** the duty in terms of Section 44 of the Finance Act (BNC)

(2h)
the assets **available for** the floating charge holders. (BNC)

(2i)
the organisms **responsible for** the sexually transmitted diseases (BNC)

The PSC has one example of sequence *the *s * for the * * *s* and 29 examples of the shortened sequence *the *s * for the * *s.* As in the BNC, most of these are embedded past participle clauses such as *The coefficients obtained for the 8 reactions.* A further set is built around nouns such as

gene, sequence post-modified by an embedded progressive participle *N + coding for*. The final pattern (6 examples) involves *responsible for* which, as we saw in the BNC examples, post-modifies a key participant (*agent, enzyme, factor, mechanism*) and introduces a biochemical process*:*

(2j)
the composition and nature of the **agents responsible for** the early metasomatic events. (PSC)

(2k)
The **enzymes responsible for** the synthesis of gramicidin S suffer degradation or inactivation. (PSC)

(2l)
Although the **mechanisms responsible for** the described effects require(s) additional studies...(PSC)

It is striking that in both patterns observed in extract 2 (the verbal group 2a-f and the nominal group 2g-l), only a very restricted set of lexical items are involved in the most closely matching sequences, in fact the same as the ones in extract 2: *directed toward(s) / responsible for*. In both cases we are dealing with conventional lexico-grammatical patterns in academic discourse: one a dynamic metaphor of topicality, roughly equivalent to 'X is of interest to': *attention* / e*ffort / interest ... is / has been directed towards*, the other a stative expression of causality, equivalent to 'X is the cause of Y': *agents / factors / mechanisms ... are responsible for*. The originality of extract 2 is that it exploits both of these phraseologies simultaneously, embedding the *responsible for* pattern within a topical introduction built around *directed toward(s).*

(JA1) *Daily Telegraph*[8]

We now turn to the analysis of initial sentences in journalistic accounts (JA) of cancer research. The first sentence of JA1reads as follows:

(3)
Scientists are a step closer to developing an early detection test and possible treatment of four of the most common and intractable cancers. (JA1)

Here is the sequence of grammatical items used in the search:

*s * a * *er to *ing an * * * and * * of * of the most * and * *s. (JA1)

The BNC has 16 examples of the sequence **sa * *er to *ing*. This corresponds to a very predictable pattern involving a verb group expressing movement or proximity (*being* / *bringing* / *taking* + *a step closer* or *a little / one step nearer*) and a complement relating to a scientific discovery (a nominalised mental process *finding, understanding*). The explicit reference to scientists in most of the examples points to journalism rather than academic science:

(3a)
SCIENTISTS believe they are **a step nearer to finding** the cause of Cot Death Syndrome, (BNC)

(3b)
but scientists are **a bit nearer to understanding** what goes on at the molecular level (BNC)

(3c)
Three papers published recently in Science move us **a little closer to understanding** the basis of the disease (BNC)

These are clearly instances of a very regular lexico-grammatical pattern used in journalism. No examples of this pattern can be found in the PSC (although some similar structures can be found, they are unrelated).

Conversely, the second part of JA1 (and in particular the sequence *of the most * and*) appears to be closer to the elaborate nominal constructions of academic writing. The PSC contains 5 examples of the sequence *of * of the most * and*. All of these are instances of the same pattern: a complex nominal in which a biochemical product or process is post-modified by a superlative epithet introduced by *most* and emphasised (sometimes redundantly) by a second epithet introduced by *and.* As can be seen, writers in the PSC have a preference for *effective* + epithet:

(3d)
a determination **of the most effective and efficient** biomarkers (PSC)

(3e)
Overview **of the most effective and convenient** reagents, (PSC)

(3f)
Acyclovir is one **of the most effective and selective** agents against herpes viruses (PSC)

(JA2) *The Independent*[9]

The first sentence of extract 4 reads as follows:

(4)
A substance found in fish oil is to be used in the treatment of cancer, after new evidence that it can shrink solid tumours and may halt the dramatic weight loss associated with the disease. (JA2)

A * *d in * * is to be *ed in the * of *, after * * that it can * * *s and may * the * * * *ed with the *. (JA2)

Here I shall only concentrate on the first main clause and the sequence: *A * * *d in * * is to be *ed in.* Although it is difficult to find the exact same sequence in the BNC, especially with three contiguous lexical items, over 50 similar examples of *a/n * found in N* can be found. This is an instance of a very regular pattern involving a complex nominal referring to a *body* or *substance,* post-modified by an embedded clause expressing a cognitive process of discovery (*encountered, found, identified*). These correspond to two related patterns: a) the finding of a body in a journalistic account, and more commonly b) a scientific 'finding' involving a specific substance. Here are examples of both from the BNC:

(4a)
The body of a boy found in the River Thames is to be re-examined to see if he was the victim of a ritual killing. (BNC)

(4b)
A potentially harmful chemical commonly found in plastic baby bottles is to be banned from their manufacture from next year. (BNC)

(4c)
A substance found in yew trees may help cancer sufferers, reports Carina Norris (BNC)

It might be thought that this pattern would also be common in science writing. However although the same structure is found in the PSC (32 examples), the pattern is not quite the same. In the PSC, the preposition *in* tends to be used to introduce nominalised processes rather than locations (as mentioned in section 2.3, in Abstracts the pattern is *found / observed in the [process X] of [participant /product Y]...*). When locations are referred to in the PSC, they are introduced by a variety of verbs expressing specific material rather than general cognitive processes, and in contexts where the

properties of biochemical products are defined as locations or roles (expressed after *in*):

(4d)
With fibronectin, **a glycoprotein deposited in** the basal membrane after debridement that produce (PSC)

(4e)
The present study demonstrates for the first time that MIF, **a cytokine involved in** the inflammatory process, is produced by human trophoblasts (PSC)

(4f)
Finally, BAL fluid contains significant levels of IL-16, **a cytokine implicated in T** lymphocyte recruitment (PSC)

Returning to extract 4, it is notable that both the extract and two BNC examples (4a-b, above) make use of the "infinitival future" (*is to be used*). I would suggest that form (combined with the passive) is more akin to the discourse of research articles than journalistic accounts. If we look in the BNC for sequences such as *is to be *ed in the * of* (the main clause in extract 4), we find examples such as the following (4g-i). It is also notable that in each case the processes expressed are closer to those of extract 4 and the PSC (*employed, implemented, used*):

(4g)
its potential **is to be employed in the evaluation of** patients with RA consistently, with close frequency, and independently of any calculating device … (BNC)

(4h)
A commercial that **is to be implemented in the teaching of** a textbook unit (BNC)

(4i)
The equipment provided in support of the AED Program **is to be used in the event of** an SCA at Gonzaga University. (BNC)

It would seem then that extract 4 is another example of a relatively "hybrid" style, with at least two lexico-grammatical patterns belonging to different types of discourse. The first main clause employs a prototypically journalistic way of presenting a "finding" (nominal post-modification of *substance*), although the core predicate in the clause expresses the "future" in a prototypically impersonal, academic manner (*is to be used*). Although

space precludes further analysis here, it is notable that the rest of the sentence involves structures which are typical of elaborate scientific prose (the embedded nominal projection *after new evidence that...* post-modification after *associated with,* etc.), but also typical of the phraseology of journalism (notably the choice of epithets: *new evidence, solid tumours, dramatic weight loss*).

(JA3) *TheGuardian*[10]

The first sentence of extract 5 reads as follows:

(5)
A substance found only in oily fish may help to fight one of the main symptoms of cancer as well as leading to new forms of treatment for some of the most resistant tumours. (JA3)

Here is the same sequence stripped of its lexical items:

A * *d only in * * may * to * one of the * *s of * as well as *ing to * *s of * for some of the most * *s. (JA3)

As with previous extracts, there is only space here to analyse one or two key features of this sentence. Perhaps the most important structure here is the post-modified subject *A * *d only in...,* which involves the same *substance found in* pattern as in extract JA3. The only difference is that the substance is found *only* in one source or location. The following BNC examples give a flavour of the general pattern:

(5a)
Studies suggest that **a substance found only in** types of manuka honey may help prevent plaque from damaging calcium phosphate in tooth ... (BNC)

(5b)
An American study has found that theaflavin-2 - **a substance found only in** black and oolong teas - was able to induce aptosis (cell death) in ... (BNC)

(5c)
The cause of this common neurological disease is thought to be a dietary factor or toxic **substance found only in** that area. (BNC)

As mentioned above, this pattern occurs in the PSC, but in a somewhat different form: the closest examples are active and passive finite clauses

such as *viral antigens were found only in the bronchial epithelium.* Also, verbs other than *found* are more likely to be used, most notably *identified* and *observed.*

The second key element of phraseology in extract JA3 is a verbal group complex introduced by *may *to * one of the * *s of.* There are no exact matches for this pattern in the BNC, although it is possible to find variations, most notably with different modal verbs or other morphological changes (such as *help + to V to / help V-ing*) or different superlatives (*one of the main / most*). The examples cited below belong to a variety of relatively formal expository genres, but it is notable that 5e and 5f are also clear examples of introductions to a "breakthrough" story. All examples share the same discourse function: the subject / theme of the clause is the key to solving or understanding some problem:

(5d)
It may **help explain one of the underlying causes** of coral decline, and is one of the most comprehensive analyses yet done on the types of viruses in a … (BNC)

(5e)
New research from the University of Adelaide **could help protect one of the world's most globally threatened tree species** - the big leaf mahogany - from ... (BNC)

(5f)
TAKE two aspirin and have a lie down used to be a doctor's cliché, but a study says it could **help fight one of the world's biggest killers** (BNC)

This phraseology is not reflected in the PSC, except for unrelated structures (*the reaction sequences ... may happen to give one or more of the products indicated*). Verb groups like *help (to) explain* and noun groups like *one of the main symptoms* can both be found in the PSC, but neither are used in the same context. The following examples from the PSC give an idea of the more usual phraseology of *may help + V* as a statement of research goals (5g-h) and the use of the superlative in defining clauses (5i-j):

(5g)
In general, genotyping **may help to categorize** the patients as mild HPA at an early stage of life. (PSC)

(5h)
The study of gene-environmental interactions **may help enhance** our understanding of how these factors cause cancer formation. (PSC)

(5i)
The release of fluoride **is one of the main advantages of** glass ionomer cements. (PSC)

(5j)
Nausea **is one of the primary symptoms** in anxiety disorders and the effect of angina on nausea may be an indirect effect via anxiety. (PSC)

(JA4) *Birmingham Post*[11]

The first sentence of extract 6 reads as follows:

(6)
Birmingham scientists believe they are on the verge of beating cancer. (JA4)

* *s * they * on the * of *ing *.(JA4)

Unlike the previous extracts (3, 4 and to a lesser extent 5), in which I find aspects of both journalistic and scientific discourse, extract 6 is almost purely journalistic. Phrases such as *scientists believe* and *beating cancer* belong prototypically to the dramatic language of a breakthrough story. Such phrases are absent of course from the PSC, where the verb *believe* is either expressed in the passive or with *we* as subject, and *beat* is a noun (*heart beat*). However, the sequence *on the * of *ing* is worth analysing in more detail. A search for this sequence results in 102 hits in the BNC, and involves a very predictable but also productive pattern of the form *(to be) on the (brink, eve, path, point, verge) of V-ing*. The discourse function of this sequence is similar to the more academic infinitival future (*is to be*) we saw in extract 5: a news story is in the process of breaking (emphasis on the "here and now") according to some higher authority or source (often expressed in a projecting clause *Scientists believe that...*) The BNC examples suggest a broad set of technical or (pseudo-)scientific topics for the breakthrough:

(6a)
He demonstrates that municipalities are **on the brink of learning** how to rezone and use other land use and devel- opment techniques that significantly reduce carbon. (BNC)

(6b)
"We are **on the eve of settling** the deal" the official, who is close to talks with Israel, told Reuters. (BNC)

(6c)
Many scientists believe that we are **on the verge of contacting** alien life-forms. (BNC)

In systemic functional terms (Halliday and Mathiessen 2004), this structure corresponds to a verbal group complex such as "to go on doing", "to keep doing", in which an initial verb (or in this case a prepositional group) expresses the "phase" or aspect of the following verb. Phase is sometimes encountered in scientific articles (*one patient, though, did go on to develop persistent ventricular arrythmias, Pain continues to be one of the more frequent causes of unplanned admission*) but this is not common, and no examples of the sequence *on the * of *ing* correspond to this pattern in the PSC (except for unrelated sequences such as *our study focused on the consequences of identifying...*).

4. Discussion

The following table sets out the basic results of the survey carried out in this study:

Extract	Patterns typical of science writing	Patterns typical of journalistic writing
RA1	... is a characteristic feature of... ... some studies suggest that nearly 90% of...	
RA2	... attention has been directed toward ... ... responsible for the complex metabolic changes ...	
JA1	... of four of the most common and intractable...	Scientists are a step closer to developing ...
JA2	… is to be used in the treatment of...	A substance found in ...
JA3		A substance found only in .... may help to fight one of the main symptoms ...
JA4		Birmingham scientists believe ... ... are on the verge of beating ...

In the analysis I have presented above, I have only been able to point out one or two sequences within each extract which I believe to be "prototypical" in either science writing or journalism. Notwithstanding the limitations of this analysis, I believe that this initial survey does show that it is possible to use grammatical items to identify extended lexico-grammatical patterns. Generally speaking, each LG pattern is exclusive to one discourse type. For example, the patterns *to be a N closer to V-ing* or *to be on the N of V-ing* in JA1 and JA4 correspond to very productive constructions which express "phase" in the verbal group. Both patterns can be seen to have very specialised discourse functions: they are both used to "break" impending news in journalistic accounts of science. In contrast, the patterns *is a Adj N of* in RA1... and *N has been directed toward(s)* in RA2 are verbal group complexes which are more typical of academic and scientific writing, and they also have their own particular discourse functions (evaluation, topicalisation etc.). It is true that some LG patterns, such as the embedded clause *a N found in* in JA2 and JA3 are found in both types of writing. However, when we look at the extended contexts of this pattern, two sub-patterns emerge: one an expression of definition in academic / science writing (*a [substance, result] [found, observed] in the [treatment] of [disease]*) and the other relating to a discovery in journalism (*a [substance] [found] (only) in [fish oil]*). This picture is complicated by the fact that in some cases (JA2 and JA3), two patterns belonging to different types of discourse can be used in the same extract. In other cases it is possible to observe a shift in discourse patterning within the same general register (RA1's move from introductory style to the present-tense reporting of results). Such instances of hybridity clearly show how variation within a single sentence is determined to a very large extent by the rhetorical functions of the surrounding text. But this does not distract from the general observation that each lexico-grammatical pattern has a distinct discourse function, and that each pattern can be broadly associated with the discourse of research or the discourse of journalism.

What are we to make of the fact that, in the majority of cases, it is in fact rather difficult to find exact matches for sequences of signs (most of the sequences analysed in the previous section involve 10 signs or less, including wild-cards and grammatical morphemes)? In lexicometrics and forensic linguistics, it has long been known that above a certain length, no two sequences of words are ever identical (for example Olsson 2004 sets this limit as low as seven identical words), and that if two long sequences do match, then there must be some degree of mutual influence (plagiarism, crypto-citation etc.). The results set out in this study appear to confirm this view. However, there are one or two factors which complicate this picture.

In particular, the viewpoint adopted in this study is somewhat different to that of the forensic linguist. Because I adopt a phraseological perspective, I have no expectation that any stretch of text will be entirely original. It is an article of faith among the so-called "contextualists" (i.e. Firth, Sinclair, Hunston and others) that the "idiom principle" guides our thoughts and words, and that much of what we can observe in discourse is based on variations of pre-established, predictable patterns of language. So the fact that many exact matches of the sequences I am looking for cannot be found is on the face of things rather surprising. However, it occurs to me that those of us who work on collocations, "fixed expressions" and other phraseological phenomena often forget the inherent creativity and variability of what we have come to think of as "formulaic" language. One reason for this complexity must be that, especially when we look at lexico-grammatical patterns above the level of the group, we are confronted with complexities at the level of the text. At this level, as we have seen in a number of examples in this study, there is often much variation in morphology and determiner use, features which in English are highly variable and sensitive to factors such as textual cohesion. The converse side of this complexity is that whenever we see even a small stretch of words from any particular text, we are almost always able to predict the text type to which it belongs. If we are able to do this fairly systematically (as shown in cloze tests, and as I have attempted to show here), then the repertoire of patterns which language users are familiar with must be very rich, and our capacity to detect and interpret such a variety of patterns must be very impressive indeed.

Notes

[1] I am thinking here of the Encode project, which aims to identify and analyse "junk DNA", or what they term "non-coding functional elements" in the human genome (Maher 2012).

[2] In the original PSC there were 150 RAs, with a total of approx. 500,000 words. Since then, I have added 250 articles to bring to total to over 1.8 million.

[3] For these data, I used the Log Likelihood metric calculated by AntConc version 3.2.4. My initial results in Gledhill (1995) were obtained using Chi-squared. Generally, Chi-squared gives more weight (= higher keyword scores) to high-frequency items, i.e. grammatical words. However, I have conducted this test using both measures, and as far as I can see, with Chi-squared, the same items occur with different scores but in the same relative order.

[4] This figure is reproduced from Gledhill (1995, 30).

[5] The reason for choosing these texts is that in 1992, while I was compiling the PSC, the research carried out by the Pharmaceutical Sciences team (under Michael Tisdale) was reported as a 'cancer breakthrough' story in over a dozen articles in

the local and national press in the UK. I have included a reference to each text used in the following sub-sections. Unfortunately, I have not been able to identify the authors of all of the journalistic texts.

[6] Tisdale, Michael. 1990. Newly identified factors that alter host metabolism in cancer cachexia. *Trends in Pharmaceutical Sciences.* 11(12): 473-475.

[7] Beck, S.A. Mulligan, H., Tisdale, M. 1990. Lipolytic factors associated with murine and human cancer cachexia. *Journal of the National Cancer Institute* 82: 1922-1926.

[8] "Cancer discovery by farmer scientis".*Daily Telegraph* 28 November 1992.

[9] Hunt, Liz. 1992. "Chemical in fish oil to be used to treat cancer". *The Independent* 30 December 1992.

[10] "Fish acid may help cancer victims". *The Guardian*, date not recorded.

[11] "Midland team may have cancer cure within year". *Birmingham Post,* 30 July 1992.

References

Anthony, Laurence. 2002. *A machine learning system for the automatic identification of text structure and application to research article abstracts in computer science*. PhD Thesis, University of Birmingham, Birmingham.

Biber, Douglas, Susan Conrad, Randi Reppen, Pat Byrd, and Marie Helt. 2002. Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly* 36(1): 9-48.

Biber, Douglas, Susan Conrad, and Viviana Cortes. 2004. "If you look at...": Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3): 371-405.

Biber, Douglas, and Federica Barbieri. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26: 263-286.

Bordet, Geneviève. 2011. *Étude contrastive derésumés de thèse dans une perspective d'anaIyse de genre*. Thèse de doctorat, 28 avril 2011. Université Paris Diderot/Paris7.

Cheng, Winnie, Chris Greaves, John McH. Sinclair, and Martin Warren. 2008. Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. *Applied Linguistics* 30(2): 236-252.

Crystal, David.[1991] 2008. *A dictionary of linguistics and phonetics* (6th Edn). London, Blackwell.

de Cock, Sylvie 1998. A recurrent word combination approach to the study of formulae in the speech of native and nonnative speakers of English. *International Journal of Corpus Linguistics* 3(1): 59-80.

Firth, John Rupert. 1957. *Papers in linguistics 1934-1951*. Oxford: Oxford University Press

Gledhill, Christopher. 1995. Collocation and genre analysis. The phraseology of grammatical items in cancer research articles and abstracts. *Zeitschrift fiir Anglistik und Amerikanistik* XLIII1(1): 11-36.

—. 2000a. *Collocations in science writing*. Tübingen: Gunter Narr.

—. 2000b. The discourse function of collocation in research article introductions. *English for Specific Purposes* 19: 115-135.

—. 2011. The lexicogrammar approach to analysing phraseology and collocation in ESP texts. *Anglais de Spécialité* 59: 5-23.

Groom, Nicholas. 2005. Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes* 4(3): 257-277.

—. 2010. Closed-class keywords and corpus-driven discourse analysis. In *Keyness in texts*, ed. Marina Bondi and Mike Scott, 59-78. Amsterdam and Philadelphia: John Benjarnins.

Halliday, Michael, and Christian Matthiessen. 2004. *An introduction to functional grammar* (3rd Edition). London: Arnold.

Hasselgren, Angela. 2002. Learner corpora and language testing: Small words as markers of learner fluency. In *Computer learner corpora, second language acquisition and foreign language teaching*, ed. Sylviane Granger, Joseph Hung and Stephanie Petch-Tyson, 143-173. Amsterdam: John Benjamins.

Hoey, Michael. 2005. *Lexical priming: A new theory of words and language*. London: Routledge.

Hunston, Susan. 2008. Starting with the small words: Patterns, lexis and semantic sequences. *International Journal of Corpus Linguistics* 13: 271-295.

Hunston, Susan, and Gill Francis.1998. Verbs observed: A corpus-driven pedagogic grammar. *Applied Linguistics* 19(1): 45-72.

—. 2000. *Pattern grammar*. Amsterdam: John Benjamins.

Hyland, Ken. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27: 4-21.

Lee, David Y.W., and Chen Xiao. 2008. Small words, big deal: Teaching the use of function words and other key items in research writing. In *Proceedings of the 8th Teaching and Language Corpora Conference*, ed. Ana Frankenberg-Garcia, Tawfiq Rkibi, Maria do Rosário Braga da Cruz, Ricardo Carvalho, Direito Cristina and Diogo Santos-Rosa, 198-206. Lisbon: Associação de Estudos e de Investigação Científica do ISLA.

Luzón Marco, María José. 1999. The phraseology and meanings of the pattern be+adjective + to-infinitive. *La Linguistique* 35(2): 47-60.

—. 2000. Collocational frameworks in medical research papers: A genre-based study. *English for Specific Purposes* 19(1): 63-86.
Maher, Brendan. 2012. Encode: The human encyclopaedia. *Nature* 489 (7414): 46-48.
Meyer, Paul. 1988. Statistical text analysis of abstracts: A pilot study on cohesion and schematicity. *Computer Corpora Des Englishen* 3: 17-40.
Olsson, John. 2004. *An introduction to language, crime and the law.* London: Continuum Books.
Renouf, Antoinette. 1992. "What do you think of that?" A Pilot study of the phraseology of the core words of English. In *New directions in English language corpora*, ed. Gerhard Leitner, 301-317. Mouton de Gruyter: Berlin.
Renouf, Antoinette, and John McH. Sinclair. 1991. Collocational frameworks in English. In *English corpus linguistics*, ed. Karin Aijmer and Bengt Altenberg, 128-143. London: Longman.
Riloff, Ellen. 1995. Little words can make a big difference for text classification. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* July 9-13 1995, 130-136. Seattle, Washington.
Scott, Mike, and Chris Tribble. 2006. *Textual patterns: Keyword and corpus analysis in language education.* Amsterdam: John Benjamins.
Simpson, Rita. 2004. Stylistic features of academic speech: The role of formulaic expressions. In *Discourse in the professions: Perspectives from corpus linguistics*, ed. Ulla Connor and Thomas A. Upton, 37-64. Amsterdam and Philadelphia: John Benjamins.
Sinclair, John McH. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.
—. 1995. *Collins Cobuild English dictionary*. (2nd Edition). London: Harper Collins.
Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics* 19(1): 143-177.
van der Wouden, Teun. 2001. Collocational behaviour in non-content words. In *Collocation: Computational extraction, analysis and exploitation. Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter*, ed. Béatrice Dalle and Geoffrey Williams, 16-23.
—. 2007. On the phraseology of stop words. *Leiden Papers in Linguistics* 4(1): 56-67.
Vergne, Jacques. 2004. Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource. In *Le poids des*

mots, Actes des 7es Journées Internationales d'Analyse Statistique des Données Textuelles, ed. Gérald Purnelle, Cédrick Fairon and Anne Dister, 1158-1165. Louvain-la-Neuve 10-12 mars 2004 / March 10-12, 2004 (JADT vol. 2).

Williams, Geoffrey C. 1998. Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics* 3(1): 151-171.