

Academic clusters: text patterning in published and postgraduate writing

Ken Hyland *Institute of Education, University of London*

An important component of fluent linguistic production is control of the multi-word expressions referred to as “clusters”, “chunks” or “bundles”. These are extended collocations which appear more frequently than expected by chance, helping to shape meanings and contributing to our sense of coherence in a text. Clusters seem to present considerable challenges to student writers struggling to make their texts both fluent and assured to readers in their new communities. This paper explores the forms, structures and functions of 4-word clusters in a corpus of research articles, doctoral dissertations and master’s theses of 3.5 million words to show not only that clusters are central to academic discourse but that they offer an important means of differentiating genres, with implications for more evidence-based instructional practices in advanced writing contexts.

Keywords: clusters, academic writing, corpus analysis, lexical patterning

Componente essenziale di una produzione linguistica scorrevole è la padronanza di espressioni multilessicali comunemente denominate *clusters*, *chunks* o *bundles*. Tali espressioni si presentano come collocazioni estese che ricorrono con frequenza superiore alla casualità, contribuendo alla formazione del significato e alla nostra percezione della coerenza testuale. I *clusters* appaiono un’area particolarmente problematica per chi, nel processo di acquisizione di una specifica scrittura disciplinare, necessita di rivolgersi alla nuova comunità scientifica con testi a un tempo scorrevoli e sicuri. Questo articolo esplora forma, struttura e funzioni dei *clusters* di quattro parole in un corpus di articoli di ricerca, tesi di dottorato e tesi di master (3.5 milioni di parole) e si propone di mostrare che i *clusters* non solo sono un elemento centrale nel discorso accademico, ma offrono un importante strumento di differenziazione dei generi, con una ricaduta verso pratiche di formazione alla scrittura avanzata sempre più basate su dati autentici.¹

Parole chiavi: *clusters*; scrittura accademica; analisi di corpora; pattern lessicali; generi di ricerca

Introduction

There are many multi-word expressions which function as structural or semantic units in English. Phrasal verbs such as *carry out* and *look after* and

idioms like *beat about the bush* and *under the weather* are fairly common in conversation and are well known to flummox even advanced second language speakers of English. Far more important than these expressions, however, particularly to academic writers, are the frequently occurring word combinations which Biber, Johansson, Leech, Conrad and Finegan (1999) call “lexical bundles” and Scott (1996) refers to as “clusters”. Essentially, these are words which follow each other more frequently than expected by chance, helping to shape text meanings and contribute to our sense of distinctiveness in a register. Thus the presence of extended collocations like *as a result of* and *it should be noted that* help identify a text as belonging to an academic register, while *in pursuance of* and *in accordance with* are likely to mark out a legal text.

These statistically linked combinations are familiar to writers and readers who frequently use a particular genre, and so come to signal competent participation in a given community of users. In contrast, the absence of such clusters might reveal the lack of fluency of a novice or newcomer to that community. Haswell (1991: 236), for example, suggests:

there can be little doubt that as writers mature they rely more and more on collocations and that the lesser use of them accounts for some characteristic behaviour of apprentice writers. Gaining control of a new register therefore requires a sensitivity to expert users’ preferences for certain sequences of words over others that might seem equally possible.

So, if learning to use the more frequent fixed phrases of a discipline can contribute to gaining a communicative competence in a field of study, there may be advantages to identifying these clusters so as to help learners acquire the specific rhetorical practices of the texts they are asked to write. In order to accomplish this, writers need a familiarity with both the clusters which characterise their disciplines and those which are valued in the particular genres of those disciplines. This study seeks to shed some light on the way language is directly experienced in academic domains by revealing how far clusters differ by genre and writing expertise, identifying the most frequent patterns in three parallel corpora of research articles, master’s dissertations and PhD theses. The study has the potential to enhance our understanding of the features of different kinds of text and to improve instruction by revealing what has a better chance of acceptance by expert readers.

An overview of clusters

The study of formulaic patterns has a long and distinguished history in applied linguistics, dating back to Jespersen (1924) and to Firth (1951), who popularised the term “collocation” along with the famous slogan: “you shall judge a word by the company it keeps” (Firth 1957). In more recent times, Nattinger and DeCarrico (1992) have emphasised the importance of frequent multi-word combinations

as a way of facilitating communication processing by making language more predictable to the hearer. Wray and Perkins (2000), for instance, argue that such sequences function as tools for social interaction and as processing short cuts, saving processing effort by being stored and retrieved whole from memory at the time of use rather than generated anew on each occasion. The extensive use of such pre-fabricated sequences as *it has been noted that* in more premeditated written genres, for instance, helps to signal the text register and reduce processing time by using familiar patterns to link new information.

Multi-word patterns, however, vary enormously in their idiomaticity and invariability, and many opaque idioms (*face the music*) and syntactic irregularities (*by and large*) fall into this category. But these are relatively rare in natural use compared with semantically transparent and formally regular clusters. With attention increasingly devoted to corpus-based analyses, many researchers (e.g. Altenberg 1998; Sinclair 1991) and syllabus designers (Lewis 1997; Willis 1990) advocate an increased pedagogic focus on what I shall call “clusters”, or recurrent strings of uninterrupted word forms. The key idea here is that of “collocation”, or “the relationship that a lexical item has with items that appear with greater than random probability in its textual context” (Hoey 2005: 3). Most clusters are structurally incomplete units, but the co-occurrence of two or more items becomes interesting if it seems to happen for a purpose and is repeated across many texts. This extension of traditional views of formulaic phrases to regular collocations such as *I’m pleased to meet you* and *I want to make three points* therefore hints at the extensive amount of formulaicity in language use, with Altenberg (1998) suggesting that as much as 80% of natural language could be patterned in this way.

The pervasiveness of this patterning has, in fact, led writers such as Sinclair (1991) and Hoey (2005) to propose radical new theories of language to replace our traditional conceptions of grammar. Instead of understanding lexical choices as constrained by the slots which grammar makes available for them, they see lexis as systematically structured through repeated patterns of use. As Sinclair (1991: 108) observes:

By far the majority of text is made of the occurrence of common words in common patterns, or in slight variants of those common patterns. 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 a text. This is totally obscured by the procedures of conventional grammar.

In other words, grammar is the output of repeated collocational groupings as words are mentally “primed” for use through our experience of them in frequent association with others (Hoey 2005). The wordings we choose are shaped by the way we regularly encounter them in similar texts.

Text receivers are therefore able to sort out what is natural from what is merely grammatical and judge whether a particular collocation “sounds right”: whether it seems usual or unusual in that context. Thus *as a result of*

is a frequent and unremarkable collocation in academic writing, but the equally possible *as resulting from* or *as an outcome of* are almost never encountered. So while clusters are simply statistical regularities of language use for the analyst, they actually reflect a lived reality for users. This, then, is a psychological association between words, and as such offers insights into the language used by writers in particular contexts. The most frequent of these strings are very common indeed and, as we can see from Table 1, vary across contexts to help us to discriminate between registers. The table shows the 25 most frequent 4-word clusters in the 4-part British National Corpus Baby edition,² each corpus consisting of about one million words and representing academic writing, imaginative writing, newspaper texts and spontaneous conversation.³

The lists illustrate several characteristics of clusters worth mentioning. The first is that they are typically building blocks of coherent discourse which span structural units. Clusters, in other words, are identified empirically purely on the basis of their frequency rather than their structure, with all those in Table 1 occurring well over 40 times per million words and the most common of them over 100 per million. Interestingly, both academic writing and conversation draw on a much larger stock of prefabricated phrases than either news or fiction, with some 800 different 4-word clusters in the conversation corpus and over 450 in the academic corpus occurring more than 10 times in one million words.

Second, it is clear that many 4-word clusters such as *one of the most* and *it is possible to* incorporate “3 word clusters in their structure” (Cortes 2004), while 4-word strings are often incorporated into longer strings. For example, *due to the fact* is part of the 5-word *due to the fact that*, which in almost all cases is part of *may be due to the fact that* and *is due to the fact that*. Four-word clusters are examined in this study because they are far more common (over 10 times more frequent than 5 words) and present a wider range of structures and functions than 3-word bundles.

Third, the four registers exhibit clear preferences for different 4-word clusters; so while there is some overlap, academic writing, for instance, shares only a few clusters with either fiction or conversation. In particular, we might note the considerable use of what Biber et al. (1999: 995) call preposition + noun phrase fragments (*on the basis of*, *in the case of*), noun phrase + *of*-phrase fragments (*a wide range of* and *one of the most*) (see also Scott and Tribble 2006: 138) as well as anticipatory *it* fragments (*it is possible to*, *it is clear that*) (Hyland and Tse 2005). Together, these comprise over 70% of 4-word patterns in academic discourse but rarely figure in conversation, where 60% of patterns are personal pronoun + lexical verb phrases (*I don't know what*, *I thought it was*) and auxiliary + active verb (*have a look at*, *do you want a*). These patterns are clearly strong register discriminators.

It seems, therefore, that control of a language involves a sensitivity to the preferences of expert users for certain sequences of words over others. Both Sinclair and Hoey take pains to point out that our different textual experiences mean that we all have a different mental concordance to draw

Table 1. Most frequent 4-word bundles from registers of the British National Corpus Baby edition

Academic	Fiction	News	Conversation
on the other hand in terms of the in the case of the end of the on the basis of as a result of the way in which it is possible to at the end of per cent of the the extent to which in the context of at the same time it is important to that there is a a wide range of it is clear that one of the most at the time of in the form of as shown in fig the rest of the can be used to in relation to the the size of the	the end of the at the end of the rest of the for the first time at the same time in the middle of the edge of the the top of the I don't want to he was going to the back of the the other side of the side of the in front of him it would have been on the edge of in front of the the middle of the what do you think a cup of tea on the other side what do you mean was going to be as if he was he shook his head	the end of the at the end of for the first time per cent of the the rest of the as a result of one of the most is one of the at the same time in the second half a member of the in the first half is likely to be by the end of will be able to the first time in the top of the in an attempt to the start of the as well as the as part of the at the start of on the other hand it would be a a spokesman for the	I don't know what no no no no do you want to I thought it was what do you want da da da da thank you very much I don't know whether have a look at are you going to do you want a you want me to what do you think I don't think so ha ha ha ha if you want to I don't want to you don't have to a bit of a know what I mean you know what I oh I don't know do you want me I don't know if or something like that

on, so that particular patterns are cumulatively loaded with the contexts we participate in. Our sense of collocational “normality” therefore depends on the genres and communities we routinely participate in, and many lexical combinations function as clear signals of competent performance.

Applied linguists and language teachers have therefore increasingly come to see clusters as important building blocks of coherent discourse and as characteristic of language use in particular settings. In academic contexts, studies by Cortes (2004) and Scott and Tribble (2006) into undergraduate and professional writing and Biber (2006) into classroom teaching and textbooks indicate the extent to which different academic discourses rely on different repertoires of lexical clusters. Various studies have also stressed the value of using collocations of apprentice texts in writing instruction, and the value of relevant models when helping students to develop control of a new genre (Granger and Tribble 1998; Hyland 2003). But despite their importance to language production, significant questions remain concerning the specific use of clusters in many key academic genres and the extent to which they may mark differences in expert and apprentice writing. Identifying the ways these genres are similar to, and different from, each other in equivalent fields may therefore place us in a better position not only to explain such differences but to employ appropriate models in the classroom.

This study therefore attempts to examine the variation of clusters across a range of disciplines to shed light on the following questions:

1. What are the most frequent 4-word clusters in these four fields of academic writing?
2. Is there evidence for systematic similarity or contrast in the forms and functions of clusters across academic genres?
3. To what extent do the clusters employed by advanced second language writers and published academics differ?
4. How might such variations be explained in terms of generic distinctiveness?

Corpora and methods

Data for the study consist of three electronic corpora of written texts. These comprised research articles, PhD dissertations and MA/MSc theses from four disciplines selected to represent a broad cross-section of academic practice: electrical engineering (EE), business studies (BS), applied linguistics (AL) and microbiology (Bio). The research article (RA) corpus is composed of 120 published papers, 30 in each of the four disciplines, and selected from leading journals recommended by expert informants and totalling 730,000 words. The PhD and master's corpora contained 20 texts in each discipline and comprised 1.9 million words and 825,000 words respectively. They were written by mainly Cantonese-speaking first language students studying at five Hong Kong universities and taught by British, American and Chinese instructors.

While the three corpora differ in terms of length, audience and purpose, they represent the key research genres of the academy, encompassing the most highly valued kinds of writing produced by students and experts. The research article is not only the principal site of disciplinary knowledge-making but, as Montgomery (1996) has it, "the master narrative of our time". One reason for this pre-eminence is the value attached to the processes of peer review as a control mechanism for transforming beliefs into knowledge, while another is the prestige attached to a genre which restructures the processes of thought and research it describes to establish a discourse for scientific fact creation. Consequently, the article is often presented to students as a model of good academic writing and as an ideal to be emulated as far as possible.

But although all research genres carry the imprint of an academic register and so are likely to be similar in many ways, the student genres present writers with different challenges and constraints. Both the PhD and master's dissertations are high-stakes genres for students, and are often the last major piece of writing they will do at university, perhaps in their lives. They carry the burden of assessment and determine future life chances, but with different expectations for particular forms of argument, cohesion and reader engagement. The problem for master's students is to demonstrate a suitable degree of intellectual autonomy while recognising readers' greater experience and knowledge of the field. For doctoral students, it is to present an understanding of disciplinary ways of working through an appropriate exposition of research and argument. Assessing the extent of similarities and differences between these genres can offer insights into apprentice and expert performance and feed into classroom practices.

The corpora were explored with WordSmith Tools (Scott 1996) using a two-step procedure. First, the lexical bundles were identified by creating a word list for each genre and generating 4-word cluster lists. Frequency of occurrence and breadth of use are the defining characteristics of these extended collocations. While Biber et al. (1999) decided to include combinations which occurred over 10 times in a million words and appeared in 5 or more texts, I decided on the more conservative cut off of 20 times per million words, and only included those which occurred in at least 10% of the texts in the sample. Second, while a corpus can tell us which clusters are frequent, an explanation of why they are frequent can only come from texts. A more qualitative analysis was therefore undertaken using a concordancer to study the textual contexts of examples and determine their functions. The frequencies and patterns were then compared across the different corpora to determine similarities and differences in the expert and student genres.

Clusters in academic writing

These criteria yielded 130 different 4-word clusters in the full corpus of 3.5 million words, totalling 12,000 individual cases or about 2% of all the words

Table 2. Structural patterns of 4-word clusters in academic writing (4 disciplines)

Structure	Examples	% of all structures	% of all cases
other prepositional phrases	on the other hand, at the same time, in the present study, with respect to the	22	27
noun phrase + <i>of</i>	the end of the, the nature of the, the beginning of the, a large number of	22	19
prepositional phrase + <i>of</i>	in the case of, at the end of, as a result of on the basis of, in the context of	15	19
noun phrases with other post-modification	the fact that the, one of the most, the extent to which, the relationship between the	12	11
passive + prepositional phrase fragment	is shown in figure, is based on the, is defined as the, can be found in	11	9
anticipatory <i>it</i> + verb/adj	it is important to, it is possible that, it is difficult to, it was found that	8	6
verb (<i>be</i>) + noun phrase	may be due to, is a matter of, is due to the, be the result of	6	5
Others	as shown in figure, should be noted that, is likely to be, as well as the	4	4
Total		100	100

in the corpus. *On the other hand* was by far the most frequent cluster, occurring 100 times per million words, and was over twice as common as the next-placed clusters, *at the same time* and *in the case of*. The top 10 clusters all occurred over 60 times per million words, and the entire list was dominated by prepositional phrase constructions and noun phrases with *of* fragments. Table 2 gives the distributions in each category in Biber et al.'s (1999: 1014–24) classification, showing examples together with the percentage of each structure in the corpus and the percentage of actual cases.

As can be seen, most of the clusters in academic writing are parts of noun or prepositional phrases and end with prepositions, articles and complimentizers such as *whether* and *that*. It is also apparent that several of these structures reflect the cautious limitations of academic discourse, typically through post-nominal modification, agent-evacuated passives and anticipatory-*it* patterns.

In addition to these structural characterisations, it is also useful to classify clusters according to their meanings in the texts (e.g. Cortes 2004). The

clusters in the corpus fall into three broad categories, which are loosely based on Halliday's (1994) linguistic macrofunctions: research, or real-world clusters, serve an ideational function, text-oriented clusters are combinations concerned with textual functions, and participant-oriented bundles express interpersonal meanings:

RESEARCH-ORIENTED. Help writers to structure their activities and experiences of the real world.

- location – indicating time and place (*at the beginning of, at the same time, in the present study*);
- procedure (*the use of the, the role of the, the purpose of the, the operation of the*);
- quantification (*the magnitude of the, a wide range of, one of the most*);
- description (*the structure of the, the size of the*);
- topic – related to the field of research (*in the Hong Kong, the currency board system*).

TEXT-ORIENTED. These clusters are concerned with the organisation of the text and the meaning of its elements as a message or argument.

- transition signals – establishing additive or contrastive links between elements (*on the other hand, in addition to the, in contrast to the*);
- resultative signals – mark inferential or causative relations between elements (*as a result of, it was found that, these results suggest that*);
- structuring signals – text-reflexive markers which organise stretches of discourse or direct reader elsewhere in text (*in the present study, in the next section, as shown in fig.*);
- framing signals – situate arguments by specifying limiting conditions (*in the case of, with respect to the, on the basis of, in the presence of, with the exception of*).

PARTICIPANT-ORIENTED. These are focused on the writer or reader of the text (Hyland 2005).

- stance features – convey the writer's attitudes and evaluations (*are likely to be, may be due to, it is possible that*);
- engagement features – address readers directly (*it should be noted that, as can be seen*).

While these classifications are sufficiently broad to minimize the possibility of overlaps between categories, no system is entirely watertight. A large sample of 2,000 cases (some 17% of all cases) were therefore checked in their context to ensure that they functioned according to the general category in which they had been placed. Only in a handful of cases was it necessary to allocate an instance to an alternative category.

I will now explore these general observations in more detail by comparing the ways clusters are used in the three genres.

Genre variations in cluster use

Analyses of the corpora of published and student texts show considerable variation in the use of 4-word clusters. The research articles contained 71 different clusters of 20 per million words or more in over 10% of texts, while the PhD theses contained 95 different clusters and the master's texts 149. Many clusters used by master's and doctoral students, therefore, are not found in the professional academic papers, or appear far less frequently. This may be because of differences in the specific topics addressed by writers in the three genres, such as the references to Hong Kong in the master's theses; but this kind of topic specificity is rare in the corpus as a whole.

A more rhetorical explanation is that the student genres are more phrasal than the published articles, and that apprentice writers are more dependent on prefabricated clusters in developing their arguments, with the PhD students closer to the expert writers. This interpretation is strengthened by the fact that not only did the writers use a greater variety of clusters, but they used them with much greater frequency. When norming figures for different text lengths, 4-word bundles meeting the criteria outlined above constituted 3.1% of the research articles, 3.8% of the PhD dissertations and 5.1% of the master's theses, indicating a considerably higher reliance on prefabricated patterns among the less experienced writers. In fact, when producing the cluster list from the master's theses I removed almost 100 4-word strings which qualified as clusters by frequency, but which occurred in less than 10% of the texts. Repetition of strings has been recognised as a problematic feature of academic texts by second language writers (e.g. Milton 1999), and here several occurred over 70 times in a single text.

Conversely, many of the clusters most frequently used in published academic writing were never, or only rarely, found in the student texts. Table 3 shows the most commonly used clusters in the three corpora in frequency order, with those items in the research article list shaded in the student lists. As can be seen, only about half of the items in the PhD and master's lists occurred in the research articles, sharing only 6 of the top 15 in the PhD texts and only 5 of the master's lists. The frequencies per million words were also often far higher in the student texts. The most used cluster, *on the other hand*, for instance, was twice as frequent in the master's theses as the articles and three times more common in the doctoral texts, with *at the same time* and *is one of the* also significantly more frequent in student texts. These strikingly higher normalised counts partly reflect the more formulaic nature of the student genres, but might equally point to students' need to display a more conciliatory approach to arguments and to demonstrate that alternative points of view have been considered.

Table 3. The 50 most frequent 4-word clusters by level (article items shared in student texts are shaded)

Research articles	No.	PhD theses	No.	Master's dissertations	No.
on the other hand	100	on the other hand	445	on the other hand	181
in the case of	94	at the same time	201	as well as the	83
on the basis of	75	in the present study	181	at the same time	80
in the presence of	60	the end of the	181	is one of the	72
at the same time	56	in the case of	177	the nature of the	68
the results of the	55	at the end of	168	in the case of	63
the extent to which	53	in terms of the	168	the results of the	62
in the context of	47	on the basis of	142	of the Hong Kong	58
as a result of	46	as well as the	133	the role of the	55
in terms of the	46	in relation to the	122	it can be seen	50
at the end of	45	is one of the	122	in the form of	49
as a function of	44	in the form of	119	the other hand the	47
it is important to	43	the fact that the	107	the performance of the	47
is shown in fig	40	at the beginning of	105	it is necessary to	46
the degree to which	40	it was found that	103	as a result of	44
the fact that the	39	to the fact that	98	as a result the	44
with respect to the	38	as shown in figure	96	can be seen that	43
as well as the	37	the nature of the	96	the relationship between the	42
the end of the	36	the relationship between the	96	in Hong Kong and	41
as shown in fig	35	with respect to the	92	the end of the	41
the magnitude of the	34	in the process of	89	at the end of	39
the effect of the	31	in the context of	88	is based on the	39
it is possible that	30	the other hand the	86	can be used to	38
the use of the	30	as a result of	85	in terms of the	37
are more likely to	29	is shown in figure	84	it is found that	37
the size of the	29	be due to the	82	as shown in fig	35
can be used to	28	can be used to	82	one of the most	35
the nature of the	27	it should be noted	81	the effectiveness of the	35
a function of the	26	was found to be	80	the result of the	35
at the beginning of	25	should be noted that	77	to ensure that the	35
in this case the	25	are more likely to	75	can be found in	34
is based on the	25	in terms of their	75	it is difficult to	33
for each of the	24	in the sense that	75	the purpose of the	33
in the absence of	24	the beginning of the	72	it should be noted	31
is likely to be	24	the results of the	72	the fact that the	31
in addition to the	23	due to the fact	67	a wide range of	30
in the form of	23	of the present study	67	is shown in figure	30
a wide range of	22	may be due to	66	on the basis of	30
can be seen in	22	one of the most	65	the accuracy of the	30
in the next section	22	is based on the	63	the structure of the	30
one of the most	22	the total number of	63	to find out the	30
the basis of the	22	can be seen as	62	with the use of	30
the beginning of the	22	in other words the	62	are summarized in table	29
a large number of	21	on the one hand	62	in addition to the	29
the other hand the	21	it can be seen	60	the operation of the	29
are shown in fig	20	it is found that	60	at the beginning of	28
the difference between the	20	for the purpose of	59	can be divided into	27
the presence of a	20	is given by equation	59	of the present study	27
the results of this	20	it is important to	59	that there is a	27
the role of the	20			the effect of the	27

Structural differences in clusters across genres

As I noted above, the most common patterns in the corpus are noun and prepositional phrases; these are common enough in the top 50 items listed here, but the three genres showed clear differences in their structural distributions. Forty-five per cent of clusters in the doctoral texts, for instance, were prepositional phrases, which also made up 9 of the top 10 clusters. These typically function, with embedded *of*-phrase fragments, to elaborate logical (particularly temporal) or textual connections between elements of an argument (1), or without the *of*-phrase, to identify a particular research or discourse context (2):

(1) Based on the market capitalization *at the end of* 1999, the United States was still the largest market in the world followed by Japan (2nd), . . . (BS PhD)

In the case of median filtering, input samples are ordered and the output is simply the middle one of the ordered samples. (Eng PhD)

(2) The activity demonstrated in the participants *in the present study* marks the worth of engaging in audience-oriented research in crisis communication in the future. (AL PhD)

The presence of these domains and motifs was reasonable *in relation to the* function of the ACV synthetase. (Bio PhD)

The research articles too contained significant *of*-phrase structures and these made up over half of all clusters in the top 50 list, but these were overwhelmingly in structures where they post-modified noun phrases. In fact, almost half of all cases of this pattern in the three corpora occurred in the articles. This is an important pattern in research writing, as it allows writers either to specify size and quantities or to highlight a feature of the research:

(3) . . . suggesting that the Cd²⁺-sensitive Ca²⁺ influx also contributes to *the magnitude of the* RNA induction. (Bio RA)

The results of the current study demonstrate that while statistically significant intercorrelations exist between these independent variables . . . (BS RA)

Note, further, *the use of the* standard variety copula *da* in the quasi-quoted speech, in contrast to the use of the regional copula *dya* in the matrix. (AL RA)

The noun phrase with *of*-phrase fragment was also the most common pattern in the top 50 clusters in the master's corpus, but these writers also made considerable use of passive patterns, employing as many of these as the other two groups combined. Biber et al. (1999: 1020) point to the relative rarity of verb phrase bundles in academic discourse; but the passive voice verb

followed by a prepositional phrase is an important means of expressing logical or locative relations, signifying graphical information and highlighting a research observation:

(4) The discussion above *is based on the* assumption the asset-pricing effects captured by size and book-to-market equity are rational. (BS MA)

The revised classification system *is shown in* Figure 3. (AL MA)

They *can be found in* a wide range of habitats, from ice fields to deserts. (Bio MS)

Interestingly, both the student genres also contained far more examples of another means of disguising authorial interpretations: the anticipatory-*it* pattern. These clusters introduced extraposed structures which foregrounded the writer's evaluation or claim without explicitly identifying its source. These phrases are adjectival or verbal, but typically point readers to how a statement should be understood:

(5) From the above table, *it can be seen that* the subjects in the experimental group had read far more books than the subjects in all the other classes in the control group. (AL MA)

It should be noted that when sample sizes are large, a significant chi-square value may reflect rather trivial differences between the predicted and sample covariance . . . (BS PhD)

Therefore, *it is necessary to* study the behaviour of the adsorption process of SPME under the influence of temperature. (Bio MSc)

Functional differences in genre clusters

Turning now to the functions of clusters, we find text-oriented strings accounting for about half of all the patterns, but with considerable inter-genre variations. Table 4 shows the percentage differences among the three categories for all 330 cluster patterns in the three corpora. While these functions can obviously be expressed in ways other than the use of 4-word clusters (e.g. Hyland 2005), the distributions point to underlying differences between the genres, and indicate something of what writers are attempting to achieve through them.

Clusters in master's theses

We can see that the master's students' discourse is characterised by a heavy use of research-oriented clusters and a relatively low use of participant-oriented

Table 4. Distribution of functional clusters by genre (%)

Genre	Research-oriented	Text-oriented	Participant-oriented	Totals
Research articles	25.5	60.3	14.2	100
PhD dissertations	34.1	54.7	11.2	100
Master's theses	48.6	42.5	8.9	100
Overall	36.1	52.5	11.4	100

forms – choices which impart a strong real-world, research-focused sense to their texts. The infrequent use of stance and engagement clusters, for example, mirrors Hong Kong students' preference for author anonymity, which is also found in their relative reluctance to employ other stance markers such as first person pronouns (Hyland 2002) and hedges (Hyland 2000). This kind of impersonality is often seen as a defining feature of expository writing by L2 students and is also the case in the Hong Kong education system, reinforced in the advice of school textbooks and teaching practices and by cultural preferences for a conciliatory, non-interventionist stance (Scollon and Scollon 1995).

Possibly more interesting, however, is the extremely high proportion of research-oriented clusters in these texts. The master's students were the only writers to refer more to their research than to its presentation, drawing particularly on those clusters which described research objects or contexts (5) and, in almost 25% of cases, those depicting procedures (6):

(5) *The structure of the resolver is similar to that of a motor.* (EE MSc)

Temperature plays *an important role in* affecting the density of oceanic environment where the chlorinity only varies to a very slight extent. (Bio MSc)

This is *the name of the* executable file, i.e. "winword", "excel", etc. (AL MA)

(6) Genre analysis was adopted to be the research methodology *to carry out the* investigation. (AL MA)

Daily spiking was required *in order to maintain* the tank mercury concentration close to the designated concentration. (Bio MSc)

Parallel processing *can be used to* carry out the multi station-runs by a number of computers in order to minimize the computation time . . . (BS MA)

This emphasis on the ways the research was conducted suggests that the real-world, physical practicalities of the investigation played a greater part in how these apprentice writers conceptualised their studies and approached

the writing task. The master's thesis is a pedagogic genre with a display and assessment purpose which clearly puts students under some pressure to showcase their ability to handle research methods appropriately and to demonstrate their familiarity with the subject content of the discipline. The additional requirement on learners to show their understanding of the conventional argument patterns of their fields seems to have been relegated to secondary importance by these writers. Perhaps this reveals the preoccupations or uncertainties of the apprentice, demonstrating competence through the control of material resources and disciplinary research practices rather than through its literacy.

Clusters in doctoral dissertations

Cluster distributions in the doctoral dissertations, in contrast, were much closer to those in the research articles, with more participant-oriented, more text-oriented, and far fewer research-oriented clusters. While participant-oriented clusters were more varied and more numerous than in the master's texts, however, the doctoral writers similarly chose to employ strings which primarily served to engage readers rather than to convey the writer's stance. These labels represent an important distinction in understanding the role and use of interpersonal resources in language, as they refer to writer- and reader-oriented features of the discourse respectively (Hyland 2005). While "stance" concerns the ways writers explicitly intrude into the discourse to convey epistemic and affective judgements, evaluations and degrees of commitment to what they say, "engagement" refers to the ways writers intervene to actively address readers as participants in the unfolding discourse.

Engagement features constituted 70% of all participant-oriented clusters in the PhD corpus, explicitly marking the presence of what Thompson (2001) calls the "reader-in-the-text", as here:

(7) *It should be noted* that the term 'system' is ambiguous between its process meaning and its structural meaning. (AL PhD)

From these tables, *it can be seen* that the best choice for the slide window is 20. (EE PhD)

Before proceeding, *it is important to* recognise that the analysis to be offered tends to be positive, and the term "optimal" used in this study is a positive one. (BS PhD)

But while we can see these forms functioning to pull readers along with the argument and guide their understanding, such a high proportion of engagement signals simultaneously represents a reluctance to adopt a more intrusive personal voice. Once again, while stance can be expressed in other

ways than 4-word clusters, the relative absence of its use in this corpus suggests that these PhD writers may be uncomfortable with explicitly aligning themselves with a particular evaluation or personally attesting to the weight they want to attribute to their claims. Such investment clearly carries a certain risk in this extremely high-stakes genre, and it appears to be a risk they do not wish to take.

On the other hand, unlike the master's students and more like the research article writers, these doctoral students used substantially more text-oriented clusters than research ones. This practice, at least in part, can also be seen as representing a more sophisticated approach to language as these advanced students sought to craft more "academic" reader-friendly prose and make more concerted attempts to engage their readers. In many cases, these 4-word clusters represent an awareness of argument and audience, and their use suggests writers' attempts to present themselves as competent academics immersed in the ideologies and practices of their fields. This is most clearly seen in the extensive use of framing devices, used to focus readers on a particular instance or to specify the conditions under which a statement can be accepted:

(8) However, *in the case of* Kodak's KIOO, which is an intricate piece of film, words are kept minimum to keep the viewer's attention. (BS PhD)

Of particular interest *in relation to the* research tasks is the analysis of topic management with respect to task content, task procedure and off-task concerns. (AL PhD)

We can see the circuit is less sensitive to the variation in the horizontal resistors *in the sense that* when the filter response is high, the power spectrum is small. (EE PhD)

These forms not only suggest a clear audience orientation and an attempt to organise their discourse in ways that readers are most likely to understand, but also lay claim to a certain disciplinary competence, demonstrating a care with both research and with language.

In addition, because the PhD corpus was over twice the length of the master's corpus, writers also found it necessary to draw on text-referential strings to structure more discursively elaborate arguments over a greater span of text. As Bunton (1999: S41) observes:

it is the very length of the research thesis which makes it all the more important for the writer to continue to orient the reader throughout the thesis as to how the current subject matter relates to the overall thesis, i.e. to maintain cohesion and coherence.

These are clusters which help organise the text by providing a frame within which new arguments can be both anchored and projected, referring to text

stages and announcing discourse goals, as in (9), or pointing to other parts of the texts to make additional material salient and available to readers in recovering the writer's intentions (10):

(9) *In an attempt to establish the research context for this inquiry, in section 2.5, I begin with the research history of language learner strategies and then present a . . .* (AL PhD)

In this section we offer evidence on the effect of corporate investment decisions on the market value of the firm. (BS PhD)

(10) When the system is in normal condition, the computer result *is shown in Figure 20* and the voltage profile of the weakest bus *is shown in Figure 21*. (EE PhD)

Their styles of being a facilitator *will be discussed in the next chapter*, indicating the favourable student factors that contributed to being a facilitator. (AL PhD)

These clusters help frame, scaffold, and present arguments as a coherently managed and organised arrangement, reflecting writers' awareness of the discursive conventions of a sustained discussion of the discoursal expectations and processing needs of a particular audience.

Clusters in research articles

The research article is clearly a very different genre from those produced by students, with a different purpose, audience and repertoire of rhetorical features. Essentially, writing for publication differs from student genres in that it is what Swales (1990) refers to as a "norm developing" practice, concerned with persuasive reporting through the review process and engagement with the professional world, rather than "norm developed", which largely displays what the student knows. The research paper is one of the main means by which academics disseminate their research and establish their reputations, exhibiting to colleagues both the relevance of their work and the novelty of their interpretations. Like dissertations, articles present an argument, but they differ in that they are broadly concerned with knowledge-making and are evaluated by peers. These differences help explain why research articles contained the most text- and participant-oriented clusters and the fewest research-oriented strings.

Not only did the research articles contain more participant-oriented clusters than the theses and dissertations, for example, but some two-thirds of these indicated the writer's stance to material rather than a reference to the reader. Although often characterised as lacking explicit appraisal and attitude, published academic writing is nevertheless clearly structured to

evoke affinity and engagement. In presenting observations, interpretations and claims, writers also project an appropriate disciplinary persona, annotating their texts to comment on the possible accuracy or credibility of a claim, the extent they want to commit themselves to it or the attitude they want to convey, as in these examples:

(11) *It is possible that* increasing the complexity and realism of the choice task may also weaken the effect. (AL RA)

Third, customers *are more likely to* form relationships with individuals and with the organizations they represent than with goods. (BS RA)

However, this *may be due to* disruption of the complex upon antibody binding, or the antibodies we have used may block the interaction. (Bio RA)

It is obvious that the partial heat resistances are provided directly by the structure function. (EE RA)

Interestingly, all stance patterns in the articles, excepting the final example in (11), functioned to withhold complete commitment to a proposition, allowing writers to present information as an opinion rather than as accredited fact. Such hedges protect the writer from possible false interpretations, and indicate the degree of confidence that it may be prudent to attribute to the accompanying statement.

A further striking contrast to the patterns in the student texts is the overwhelming preponderance of text-oriented clusters in the research article corpus. This is the most discursively crafted and rhetorically machined genre of the three, with almost two-thirds of its clusters presenting the research to a disciplinary audience by engaging with a literature, providing warrants, establishing background, connecting ideas, directing readers around the text and specifying limitations. Clearly, results and interpretations need to be presented in ways that readers are likely to find persuasive, and so writers must draw on these to express their positions, represent themselves and engage their audiences. Perhaps not surprisingly, and as in the PhD texts, framing clusters were most common in this corpus, as they help writers to elaborate arguments by highlighting connections, specifying cases and pointing to limitations; but we also find significant numbers of resultative markers in the articles, 225 cases in all, pointing readers to the writer's interpretations and understandings of research processes and outcomes.

This is a key function in research writing as these clusters signal the main conclusions to be drawn from the study and highlight the inferences the writer wants readers to draw from the discussion:

(12) *The results of the* mating experiments clearly indicate the existence of two ISGs in *C. subnuda*. (Bio RA)

On the theoretical level, our *results suggest that the* perspective of opportunism may not axiomatically hold in all asymmetric contexts. (BS RA)

As a result the combination of nonuniform doping in the emitter and base could maximize the cut-off frequency and maximum oscillation frequency of the heterojunction bipolar transistor. (EE RA)

As we can see from the first example, resultative markers can frame an assertive construal of events, boosting the writer's position and directing readers to a categorical understanding. More often, however, they preceded a more conciliatory stance, downplaying any confidence the writer might have in his or her interpretation and opening a discursive space in which the reader might feel free to dispute it.

Conclusions

Phraseology has been one of the most rapidly growing areas in applied linguistics over the past 25 years, revealing that routine strings of words are pervasive in natural language use. My main purpose in this study was to explore the extent to which phraseology contributes to academic writing by identify the most frequent 4-word clusters in three key academic genres and elaborating their structures and functions. The findings support earlier studies by Cortes (2004) and Scott and Tribble (2006) which show considerable variations in the frequency of forms, structures and functions across student and expert writing. Clusters, in other words, should not only be regarded as a basic linguistic construct; their distribution can also be an effective way of characterising genres within a single register.

This study indicates that professional academics, doctoral students and master's level students draw on different resources to develop their arguments, establish their credibility and persuade their readers. The research articles, for instance, contained far fewer clusters and far fewer *different* clusters overall; they included largely different clusters to the student genres, with less than half of the forms overlapping in the most common 50 items, and with far more *noun phrase + of* structures; and they revealed more participant strings and included a far higher proportion of text-oriented clusters. Clusters in the master's texts displayed diametrically opposite patterns.

The fact that the three genres are characterised by different cluster patterns is no accident, nor does it necessarily reflect deficiencies in the English used by these second language writers or in their ability to control the conventions of academic writing in a foreign language. All the texts examined were judged successful by expert readers and were awarded high passes. That master's students made the greatest use of clusters in their

writing and the professional writers the least could suggest a greater reliance on formulaic expressions by less confident or proficient students in constructing their texts. We might, however, also see these patterns as important ways of shaping texts for different purposes and readers. The master's thesis, for example, is essentially a pedagogic genre; and while its writers share a persuasive goal with doctoral and professional writers, the burden of assessment at this level puts greater emphasis on a discourse which displays the writer's research skills and a practical disciplinary competence. An emphasis on research features in these circumstances might therefore be an appropriate persuasive strategy rather than a departure from the norms of research writing. Such speculations require further research to establish the extent of variations across a broader range of texts, but it is clear that clusters play a key role in the main genres of the academy at different levels of writer experience and expertise.

The findings also have clear implications for pedagogic practice. First, evidence from learner corpora help improve descriptions of the target language and provide more realistic models for students. It alerts us to the need to understand the kinds of text our students need to write rather than rely on the massive literature which describes the research article. Second, an improved understanding of learner output illuminates all aspects of pedagogy from tasks to curriculum (Granger 2002). While frequency should never, by itself, determine classroom decisions, learner corpus data can play an important role in the selection, sequencing and structuring of teaching content. Third, teaching materials can benefit from the findings of cluster research in different genres, allowing teachers to focus on the specific ways of creating meanings appropriate to particular kinds of writing. Finally, the use of relevant genre data refocuses instruction on form, and can provide a basis for methodological practices which involve data-driven learning. While the use of learner corpus data in the classroom is somewhat controversial, research suggests that it can offer an important contribution to learning in advanced contexts (Granger, Hung and Petch-Tyson 2002).

In sum, I suggest that the study of clusters offers insights into a crucial, and often overlooked, dimension of language use, providing a better understanding of the ways writers employ the resources of English in different contexts, and with the potential to inform advanced academic literacy instruction.

Notes

1. Thanks to Marina Bondi for a poetic translation.
2. Details of this corpus can be found at <http://www.natcorp.ox.ac.uk/corpus/index.xml.ID=products>
3. Strings of numbers, which are among the most frequent conversational clusters, have been deleted from the list.

References

- Altenberg, B. (1998) On the phraseology of spoken English: the evidence of recurrent word combinations. In A. Cowie (ed.), *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press. 101–22.
- Biber, D. (2006) *University Language: A Corpus-Based Study of Spoken and Written Registers*. Amsterdam: Benjamins.
- S. Johansson, G. Leech, S. Conrad and E. Finegan (1999) *Longman Grammar of Spoken and Written English*. Harlow: Pearson.
- Bunton, D. (1999) The use of higher level metatext in PhD theses. *English for Specific Purposes* 18: S41–S56.
- Cortes, V. (2004) Lexical bundles in published and student disciplinary writing: examples from history and biology. *English for Specific Purposes* 23: 397–423.
- Firth, J.R. (1951) Modes of meaning. *Essays and Studies* (English Association): 118–49.
- (1957) *Papers in Linguistics*. London: Oxford University Press.
- Granger, S. (2002). A bird's-eye view of learner corpus research. In Granger, Hung and Petch-Tyson (2002: 3–36).
- J. Hung and S. Petch-Tyson (eds.) (2002) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: Benjamins.
- and C. Tribble (1998) Exploiting learner corpus data in the classroom: form focused instruction and data-driven learning. In S. Granger (ed.), *Learner Language on Computer*. Harlow: Longman.
- Halliday, M.A.K. (1994) *Functions of language*. 2nd edn. London: Arnold.
- Haswell, R. (1991) *Gaining Ground in College Writing: Tales of Development and Interpretation*. Dallas: Southern Methodist University Press.
- Hoey, M. (2005) *Lexical Priming: A New Theory of Words and Language*. London: Routledge.
- Hyland, K. (2000) 'It might be suggested that...': academic hedging and student writing. *Australian Review of Applied Linguistics* 16: 83–97.
- (2002) Authority and invisibility: authorial identity in academic writing. *Journal of Pragmatics* 34.8: 1091–1112
- (2003) Genre-based pedagogies: a social response to process. *Journal of Second Language Writing* 12.1: 17–29.
- (2005) Stance and engagement: a model of interaction in academic discourse. *Discourse Studies* 7.2: 173–91.
- and P. Tse (2005) Hooking the reader: a corpus study of *evaluative that* in abstracts. *English for Specific Purposes* 24: 123–39.
- Jespersen, O. (1924) *The Philosophy of Grammar*. London: Allen & Unwin.
- Lewis, M. (1997) *Implementing the Lexical Approach*. Hove: Language Teaching Publications.
- Milton, J. (1999) Lexical thickets and electronic gateways: making text accessible by novice writers. In C.N. Candlin and K. Hyland (eds.), *Writing: Texts, Processes and Practices*. London: Longman. 221–43.
- Montgomery, S. (1996) *The Scientific Voice*. New York: Guilford Press.
- Nattinger, J. and J. DeCarrico (1992) *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.
- Scollon, R. and S. Scollon (1995) *Intercultural Communication*. Oxford: Blackwell.
- Scott, M. (1996) *Wordsmith Tools 4*. Oxford: Oxford University Press.

- and C. Tribble (2006) *Textual Patterns*. Amsterdam: Benjamins.
- Sinclair, J. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Swales, J. (1990) *Genre Analysis*. Cambridge: Cambridge University Press.
- Thompson, G. (2001) Interaction in academic writing: learning to argue with the reader. *Applied Linguistics* 22.1: 58–78.
- Willis, D. (1990) *The Lexical Syllabus*. London: Collins.
- Wray, A. and M. Perkins (2000). The functions of formulaic language. *Language and Communication* 20: 1–28.

e-mail: k.hyland@joe.ac.uk

[Received August 12, 2007]