UNIVERSITY OF MURCIA

International Journal of English Studies

IJES

www.um.es/ijes

Academic Lexis and Disciplinary Practice: Corpus Evidence for Specificity

KEN HYLAND*
POLLY TSE*
University of London

ABSTRACT

The presence of unfamiliar words and expressions in academic texts is a serious obstacle to students reading in a second language. EAP has responded to this challenge by taking the view that there is a common core of academic vocabulary which is frequent across an academic register. This paper briefly considers this view by examining the range, frequency, collocation, and meaning of items on *the Academic Word List* (AWL) in a large multidisciplinary corpus. Our corpus analysis shows that individual lexical items on the list often occur and behave in different ways across disciplines and that words commonly contribute to 'lexical bundles' which also reflect disciplinary preferences. Our findings question the widely held assumption that there is a single core vocabulary needed for academic study and suggests that teachers should assist students towards developing a more restricted, disciplinary-based lexical repertoire.

KEYWORDS: Academic Word List, vocabulary, lexical bundles, disciplinary writing, specificity

Address for correspondence: Prof Ken Hyland. School of Culture, Language and Communication, Institute of Education, University of London, 20 Bedford Way, London WC1H 0AL. Tel: 020 7612 6789; e-mail: k.hyland@ioe.ac.uk

Polly Tse. Language Centre, Hong Kong University of Science and Technology, Clearwater Bay Road, Hong Kong; e-mail: lcpolly@ust.hk

I. READING AND ACADEMIC LEXIS

Reading in English is consistently shown to be of great concern to Non-Native English speaking students at tertiary level (*e.g.* Hyland, 1997; Littlewood & Liu, 1996) and understanding previously unencountered 'technical' vocabulary and 'difficult words' appears to cause the greatest trouble (Evans & Green, 2007). A key component of successful language learning is therefore control of the routine patterns of expression (Wray, 2000) and "semi-technical vocabulary" (Farrell, 1990) which students encounter in their disciplinary reading.

The response of materials writers and curriculum developers working in English for Academic Purposes (EAP) has largely involved 'register targeting' by seeking to identify lexical items which are reasonably frequent in a wide range of academic genres but are relatively uncommon in other kinds of texts (Coxhead & Nation, 2001). This vocabulary is seen as contributing an important element to an 'academic style' of writing and being 'more advanced' (Jordan, 1998) than the core 2,000 to 3,000 words that typically comprise around 80% of the words students are likely to encounter in reading English at university (Carter, 1998; Nation, 1990). Vocabulary, in other words, is typically seen as falling into three main groups (Nation, 2001):

1. High frequency words such as those included in West's (1953) General Service List of the most widely useful 2,000 word families in English, providing coverage of about 80% of most texts.
2. An academic vocabulary of words which are reasonably frequent in academic writing and comprise some 8% to 10% of running words of academic texts.
3. A technical vocabulary which differs by subject area and covers up to 5% of texts.

First year undergraduate students are said to find academic vocabulary a particularly challenging aspect of their learning (Li & Pemberton, 1994) because, unlike technical vocabulary, it serves a largely supportive role and items are "not likely to be glossed by the content teacher" (Flowerdew, 1993: 236). Many of these words also occur too infrequently to allow incidental learning (Worthington & Nation, 1996), encouraging researchers and teachers to develop vocabulary lists for the direct teaching of these terms. Teachers have been assisted here by the findings of corpus-based inventories, the most widely used being

the *Academic Word List* (AWL) (Coxhead, 2000; Coxhead & Nation, 2001). This contains 570 word families (the base word plus its inflected forms and transparent derivations) seen as essential for students irrespective of their chosen field of specialization. The 3,112 individual items in this inventory do not occur in West's general service list and were fairly frequent in a corpus of 3.5 million words of academic genres and across a range of disciplines in the arts, commerce, law, and sciences (Coxhead, 2000:221).

There is no doubt that the AWL is an impressive undertaking, representing the most extensive investigation into core academic vocabulary to date and now widely used in teaching materials (e.g. Schmitt & Schmitt, 2005). It remains unclear, however, how far it can be said to represent the lexical composition of academic writing in English. The notion that some words occur more frequently in academic texts than in other domains is uncontroversial and seems to fit well with EAP's distinctive approach to language teaching, based on identifying and teaching features specific to the particular disciplinary needs of learners. But while this general academic vocabulary might seem to offer good learning returns with less investment of time and effort, the view that students should be developing a *general* academic vocabulary is actually quite contentious.

It is by no means certain that there is a single literacy which university students need to acquire to participate in academic environments and we believe that a perspective which seeks to identify and teach such a vocabulary fails to engage with current conceptions of literacy and EAP, ignores important differences in the collocational and semantic behavior of words, and does not correspond with the ways language is actually used in academic writing. It is, in other words, an assumption which could seriously mislead students. In this paper we explore this view and offer some evidence for the disciplinary specific nature of lexis.

II. LEXICAL SPECIFICITY: EXPLORING THE AWL

To explore how effective the items on the AWL might be for students in different fields, we compiled a corpus of academic writing in eight disciplines representing the sciences, engineering, and the social sciences. The corpus comprised research articles, textbook chapters, science squibs, and academic book reviews which students might be

expected to read at university as well as doctoral theses, masters dissertations and undergraduate project reports, 620 texts in all totaling 3.3 million words.

Using RANGE, a program developed by Nation (2002) and used to create the AWL, we found all 570 of the AWL word families occurred in our corpus, with 541 occurring in all three fields. The AWL covered 10.6% of the words in the corpus and provided an accumulative coverage of 85% when added to the 2,000 words of the General Service List, representing roughly one unknown word in every seven words of text (Hyland & Tse, 2007). But while the list offered a good *overall* coverage, items were not evenly distributed across the entire corpus. Students in the sciences, for example, are not well served by the list, suggesting that writing in the sciences demands a more specialized and technical vocabulary, but as we shall discuss below, the fact that all disciplines shape words for their own uses seriously undermines attempts to construct a 'core' academic vocabulary.

Defining items as *frequent* only if they occurred above the mean for all AWL items in the corpus (i.e. 597), we found only 192 families, or about a third of the AWL items, met this criteria. The research terms *process, analyze, research, data* and *method* were the most common while *commence, concurrent, levy* and *forthcoming* were among 23 extremely infrequent families, occurring less than 60 times in the corpus (below 10% of the overall mean). Moreover, it appears that some items are frequent overall because of their concentration in one or two fields. 15 of our top 50 items, for example, had over 70% of their occurrences in one field. Taking the means of individual fields as a benchmark, we found that of the 192 families which were frequent overall, only 82 were frequent in all three fields and 50 in just one. Nor were the same items the most frequent in all fields. Table 1 shows that only *analyze* and *process* of the overall most frequent items also occurred in the top ten most frequent families in each field.

Overall (3 Fields)			***Engineering***			***Sciences***			***Social Sciences***		
Family	**Freq**	**%**	**Family**	**Freq**	**%**	**Family**	**Freq**	**%**	**Family**	**Freq**	**%**
Process	4501	1.3	Equate	1418		Data	1395	1.8	Research	3261	1.6
Analyze	4498	1.3	*Process*	1143		Method	1271	1.6	Strategy	2795	1.4
Research	3841	1.1	Design	999		*Process*	1118	1.4	Culture	2583	1.3
Data	3789	1.1	Method	920		*Analyze*	1029	1.3	*Analyze*	2574	1.3
Method	3214	0.9	Data	913		Concentrate	865	1.1	*Process*	2240	1.1
Vary	3156	0.9	*Analyze*	895		Require	848	1.1	Consume	1947	1.0
Strategy	3001	0.9	Function	847		Function	759	1.0	Response	1910	1.0
Culture	2962	0.9	Require	844		Obtain	750	1.0	Individual	1894	0.9
Function	2909	0.9	Output	839		Extract	739	0.9	Participate	1800	0.9
Significant	2742	0.8	Input	818		Similar	726	0.9	Significant	1762	0.9

Table 1: Most frequent items by field with percentages of families in that field

Distributions are also unequal when we looked at the least frequent words. Using 10% of the mean in each field as a reference, we found 78 families to be extremely infrequent in one field, 63 in two fields and 6 in all three. In other words, 27% of all the AWL families have a very low occurrence in at least one field and so have an extremely low chance of being encountered by students.

Comparing the occurrence of words relative to the mean helps to determine the relative significance of particular words in different fields, but a more accurate picture is obtained by norming frequencies to overcome variations in the sizes of sub-corpora. Theoretically, an even distribution would be about 33% of each item in each of the three fields, but no family met this criteria and over half of all items occurred mainly in one field only. Of the 570 AWL families, 534 (94%) have irregular distributions across the three fields with 40% of items having at least 60% of all instances in just one field. Among the most frequent items, over 90% of all cases of *participate, communicate, output, attitude, conflict, authority, perspective and simulate* occurred in one field. In fact, only 36 word families were relatively evenly distributed across the science, engineering and social science fields, and so might therefore qualify for an academic word list. Of these, however, only 22 might be considered as *frequent* by our criterion, and only seven were in the top 60

items. Just six families appeared in the top 60 of both Coxhead's list and our own: *analyze, consist, factor, indicate, period* and *structure*.

This concentration of items is also apparent when we look at distributions *within* fields. Table 2 shows 283 items in engineering (52% of all families) having over 65% of all cases in just one discipline, 244 items in the sciences (43%) with over 65% in just one discipline, and 128 (22.5%) of items in the social sciences with over 65% in one discipline. Overall, only one family occurred roughly equally across the three disciplines in the sciences and seven in the social sciences although engineering seems to be easier to identify a common semi-technical vocabulary with 47 items appearing equally across electrical and mechanical engineering.

Disciplines	Families	Total 40-64%	Total of all items occurring in one discipline 65-79%	Over 80%
Engineering	542	259(47.8%)	133(24.5%)	150 (27.7%)
Sciences	568	322(56.7%)	116(20.5%)	128 (22.5%)
Social Sciences	570	409(71.8%)	74(13.0%)	54 (9.5%)
Overall	570	336(59.0%)	110(19.4%)	114 (20.0%)

Table 2: Concentration of items in disciplines (% adjusted for corpus size)

Once again then, the patterns point to a more complex picture of language use in the disciplines than notions of a general academic vocabulary allow, pointing to more specialized language uses.

III. MEANINGS AND USES OF WORDS

There is a further difficulty with compiling a 'common core' of academic vocabulary as items also *behave* differently across disciplines. Most words have more than one sense yet students need to be confident that they are understanding words in the right way when reading academic texts. This means that a vocabulary list must either avoid items with clearly different meanings and dissimilar co-occurrence patterns, or these must be taught separately rather than as parts of families. We must, then, be cautious about claiming generality for

families whose meanings and collocational environments may differ across each inflected and derived word form (Oakey, 2003).

Wang and Nation (2004) explored this possibility in the AWL and found only a small number of families which contained homographs, or unrelated meanings of the same written form and suggested that words have essentially similar meanings across fields. In our corpus, however, there were clear preferences for particular meanings and collocations in different disciplines. As brief examples, we might take the two most frequent AWL items in the corpus, *process* and *analyze*, both of which occur far more often in academic discourse than in other registers.

Despite its high frequency in all three fields, the word *process* is far more likely to be encountered as a noun by science and engineering students than by social scientists. This is the result of nominalization (Halliday, 1998), which refers to the way that writers in the sciences regularly transform experiences into abstractions to create new conceptual objects. Embedding an item such as *process* into complex abstract nominal groups produces terms such as:

- *A constant volume combustion process…*
- *the trouble call handling process…*
- *processing dependent saturation junction factors…*
- *the graphical process configuration editor…*

Such constructions allows writers to give new objects stable names and to manage the information flow in a text more efficiently, but they do not help novices to unpack specialized meanings from the individual lexical item. We believe this is therefore likely to present difficulties to both native and non-native English speaking students.

Similarly, *analyze* seems to be used differently across fields, occurring regularly as a noun in the social sciences but with engineering students six times more likely to come across the form *analytical*. There are also semantic differences. The word *analysis*, for instance, tends to be associated with particular types of approach, so that it appears in disciplinary specific compound nouns such as *genre analysis* or *neutron activation analysis*. The verb form also has field-specific meanings, with scientific uses referring to

methods of determining the composition of a substance (1), while in the social sciences it has a sense closer to considering something carefully (2):

(1) In order to *analyze* the activity of the somatostatin promoter in DTC1 cells after integrating into the host chromosome, two pools of DTC1 stable transfectants … (Bio PhD)

We *analyze* the MSHG image of two neighboring domains and two parts A and B of a domain wall … (Phy RA)

(2) That opportunity lies about 10 years into the future for this sample, when we can *analyze* cumulative conviction records from age 14 to age 30 to span the desistance process … (Socio RA)

This paper attempts to analyze whether the currency attack on Hong Kong dollar since the outbreak of Asian financial crisis was …(Bus MA)

In fact, analysis of potential homographs in the AWL reveals a considerable amount of semantic variation across fields. Table 3 shows the main meanings for selected words with different overall frequencies in the AWL together with their distributions.

The table shows that even where items are very frequent, there are still wide variations in preferred uses, with social science students far more likely to meet *consist* as meaning 'to stay the same' and science and engineering students very unlikely to come across *volume* as a book. With less frequent words the preferred meanings differ dramatically. More worrying, these preferred uses become even more apparent when we consider patterns at the disciplinary level (Hyland & Tse, 2007).

Family	**meaning**	**Science**	**Engineering**	**Social Science**
Consist	stay the same	34	15	55
(rank 41)	made up of	66	75	45
Issue	flow out	7	6	18
(46)	topic	93	94	82
Attribute	feature	83	35	60
(93)	ascribe to	17	65	40
Volume	book	1	7	50
(148)	quantity	99	93	50
Generation	growth stage	2	2	36
(245)	Create	98	98	64
Credit	acknowledge	0	60	52
(320)	payment	100	40	48
Abstract	précis/extract	76	100	13
(461)	Theoretical	14	0	87
offset	counter	0	14	100
(547)	out of line	100	86	0

Table 3: Distribution of meanings of selected AWL word families across fields (%)

The fact that words take on additional meanings as a result of their regular co-occurrence with other items may also create difficulties for learners working from a general academic wordlist. The term *value* in computer science, for instance, is often found as *value stream* (21% of all cases) and *multiple-value attribute mapping* (7% of all cases). Even high frequency items such as *strategy* have preferred associations with *marketing strategy* forming 11% of all cases in business, *learning strategy* making up 9% of cases in applied linguistics, and *coping strategy* comprising 31% of cases in sociology.

In sum, these different word choices, collocates and fixed phrases colour the everyday uses of words with more particular discipline-specific meanings, reflecting how writers need to represent themselves and their ideas through a locally appropriate theoretical and methodological framework.

IV. DISCIPLINARY SPECIFIC BUNDLES

The disciplinary specific patterns we have found in the uses of individual words are also apparent in the distribution of 'lexical bundles', or strings of words which follow each other more frequently than expected by chance. Such stable word combinations are an important part of a discipline's discoursal resources but enormously complicate the business of constructing general word lists. By breaking into single words items which may be better learnt as wholes, vocabulary lists simultaneously misrepresent disciplinary specific meanings and mislead students.

Bundles, in fact, are familiar to writers and readers who regularly participate in a particular discourse. The very 'naturalness' of extended collocations like *as a result of, it should be noted that,* and *as can be seen,* for example, signal competent participation in an academic register and (Biber, 2006; Biber, Conrad & Cortes, 2004; Scott & Tribble, 2006). Wray and Perkins (2000), for instance, argue that such sequences function as processing short-cuts by being stored and retrieved whole from memory at the time of use rather than generated anew on each occasion. Text receivers are therefore able to sort out what is natural from what is merely grammatical and judge whether a particular collocation 'sounds right' in that context. In fact, it is often a failure to use native-like formulaic sequences which identifies students as outsiders and there is a general consensus that formulaic sequences are difficult for L2 learners to acquire (*e.g.* Yorio, 1989).

All this has led Sinclair (1991) and Hoey (2005) to propose that lexis is systematically structured through repeated patterns of use, rather than simply filling the slots which grammar make available for it. As Sinclair (1991, p. 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.

In other words, grammar is the output of repeated collocational groupings. Sentences are typically made up of interlocking bundles as everything we know about a word is a result of our routine encounters with it, so that when we formulate what we want to say, the wordings we choose are shaped by the way we regularly find them in similar texts. In academic contexts this means that bundles not only help identify communicative practices in particular disciplines, but help define the disciplinary texts themselves.

Examining a 3.5 million word corpus of research articles, PhD theses and Masters dissertations in four disciplines, the first author found 240 different 4-word bundles, totalling nearly 16,000 individual cases (Hyland, 2008). *On the other hand* was by far the most frequent of these, occurring about 200 times per million words, and was over twice as common as the next placed bundles, *at the same time* and *in the case of*. There was, however, considerable variation in disciplinary preferences, with Electrical engineering containing the greatest range of different 4-word bundles and Biology the fewest. This greater reliance on prefabricated structures could be a consequence of the relatively abstract and graphical nature of technical communication where arguments are often based on findings presented in visual form with formulaic links between them.

There were also considerable differences in the 4-word bundles themselves across disciplines. Table 4 shows the fifty most commonly used bundles in the four fields in frequency order, with items occurring in all four disciplines marked in bold and those occurring in three disciplines shaded.

Biology	Electrical Eng	Applied Ling	Business Studies
in the presence of	**on the other hand**	**on the other hand**	**on the other hand**
in the present study	as shown in figure	**at the same time**	**in the case of**
on the other hand	**in the case of**	in terms of the	**at the same time**
the end of the	is shown in figure	on the basis of	at the end of
is one of the	it can be seen	in relation to the	on the basis of
at the end of	as shown in fig	**in the case of**	**as well as the**
it was found that	is shown in fig	in the present study	the extent to which
at the beginning of	can be seen that	the end of the	at the end of the
as well as the	can be used to	the nature of the	significantly different from zero
as a result of	the performance of the	in the form of	are more likely to
it is possible that	as a function of	**as well as the**	the relationship between the
are shown in figure	is based on the	at the end of	**the results of the**
was found to be	with respect to the	the fact that the	the hang seng index
be due to the	is given by equation	in the context of	on the other hand the
in the case of	the effect of the	is one of the	in the context of
is shown in figure	the magnitude of the	in the process of	as a result of
the beginning of the	**at the same time**	**the results of the**	the performance of the
the nature of the	in this case the	in terms of their	hong kong stock market
the fact that the	it is found that	to the fact that	is positively related to
may be due to	the size of the	in the sense that	are significantly different from
are summarized in table	be seen that the	the relationship between	in terms of the
has been shown to	the accuracy of the	the	the degree to which
an important role in	**as well as the**	of the hong kong	in the long run
at room temperature for	the same as the	at the beginning of	in the united states
at the same time	is one of the	the role of the	the nature of the
can be used to	a function of the	of the present study	the total number of
in the absence of	as a result the	as a result of	the size of the
as shown in figure	**the results of the**	one of the most	in the number of
with respect to the	in the form of	can be seen as	it is important to
used in this study	is assumed to be	it is important to	the standard deviation of
was added to the	of the power system	it should be noted	in the hong kong
a result of the	it is necessary to	on the one hand	with respect to the
in addition to the	it is possible to	can be found in	of the number of
the quality of the	the length of the	the ways in which	in the form of
are listed in table	are shown in fig	in other words the	the difference between the
is due to the	can be obtained by	on the other hand the	by the end of
the presence of a	in terms of the	the starting point of	the effect of the
the results of the	are shown in figure	be seen as a	is consistent with the
was found in the	is due to the	in the eyes of	the quality of the
were found to be	the structure of the	the beginning of the	as a result the
a wide range of	is defined as the	should be noted that	can be used to
the effect of the	it was found that	that there is a	in addition to the
in the presence of the	on the other hand the	at the level of	standard deviation of the
to the presence of	the presence of the	for the purpose of	the fact that the
was used as a	with the use of	in hong kong and	in the presence of
as a result the	is the same as	are more likely to	we assume that the
have been shown to	it can be observed	the meaning of the	is more likely to
in this study the	it is because the	on the part of	the efficiency of the
is possible that the	than that of the	the purpose of the	the price of the
the base of the	will be discussed in	a wide range of	a wide range of
		the use of the	

Table 4: Most frequent 50 4-word bundles in four disciplines (Hyland, 2008)

It can be seen that over half the items in each list do not occur at all in any other discipline and only 30% of the strings in each discipline are found in two other fields. The discipline-specificity of these preferences for 4-word bundles is illustrated by the bold and shaded items, with only five bundles shared across all four disciplines and just 14 bundles occurring in three disciplines. Electronic engineering and Applied Linguistics shared just nine bundles, for example. The best candidate bundles for a general list are *on the other hand, in the case of, as well as the,* and *the end of the,* all of which occur in the top band of bundles in at least three disciplines and so comprise bundles with high frequencies across fields.

Unsurprisingly, the greatest similarities are between broadly cognate fields. Business Studies and Applied Linguistics share 18 items and Biology and Electrical Engineering have 16 bundles in common with *it was found that, is shown in figure, as shown in figure, is due to the,* and *the presence of the* not found in the social science lists. The contrasts between these two short lists reflect something of the argument patterns in the two domains, with those in the first group largely connecting aspects of argument and those in the second group avoiding authorial presence while pointing to graphs and findings. It is worth noting that while there were no bundles referring to tables or figures in the applied linguistics corpus and only two in the business texts, both science lists included these as among their most frequent strings.

V. THE FUNCTIONS OF BUNDLES

While it is useful to consider the lexical composition of formulaic strings, understanding their functional distributions is a key way in which teachers can help their students with reading assignments. The bundles in this corpus can be classified into the three categories of research, text and participants (Hyland, 2008):

Research-oriented - help writers to structure their activities and experiences of the real world.

- **Location** - indicating time/place (*at the start 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, the surface of the)*
- **topic** - related to the field of research (*in the Hong Kong, the currency board system).*

Text-oriented – concerned with the organisation of the text and its meaning 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, as can be seen*)

VI. DISTRIBUTION OF BUNDLE FUNCTIONS

Analysing the corpus reveals substantial disciplinary differences, pointing to variations in what writers are attempting to achieve through their linguistic choices. Table 5 indicates the principal differences.

Discipline	Research-oriented	Text-oriented	Participant-oriented	Totals
Biology	48.1	43.5	8.4	100
Electrical Eng	49.4	40.4	9.2	100
Applied Linguistics	31.2	49.5	18.6	100
Business Studies	36.0	48.4	16.6	100
Overall	41.2	45.5	13.2	100

Table 5: Distribution of bundle functions by discipline (%)

One obvious difference is the heavier use of research-oriented bundles in the science and engineering texts, a preference which amounted to almost half of all bundles in the science/technology corpora. The overall effect of this use is to convey a greater real-world,

laboratory-focused sense to writing in the hard sciences which, in turn, plays an important role in conveying the grounded, experimental basis of research in the hard sciences. While many of these bundles specify models, equipment, materials or aspects of the research environment (3), almost half of all cases depicted research procedures, showing the ways that experiments and research were conducted (4):

(3) the input terminal of the operational amplifier is determined by the potentiometer setting of the resistor R2 Which thus controls *the slope of the* segment that is being simulated. (EE RA)

The depth of the leaf litter layer at *the base of the* reedbeds was measured by a meter rule (0-30cm) … (Bio MSc)

(4) A programmable gain amplifier *can be used* to improved the dynamic range of the inputs … (EE RA)

Then sample buffer *was added to the* pellet which was boiled for 10 minutes followed by transfer of the sample buffer-protein… (Bio MSc)

New knowledge in these disciplines is presented and accepted on the basis of empirical demonstration designed to test hypotheses related to gaps in knowledge. The rhetorical conventions of the field, help contribute to this epistemological framework and the presence of these patterns of 4-word bundles is likely to be a key feature for students reading in the sciences.

The Applied Linguistics and Business Studies corpora, in contrast, were dominated by text-oriented strings reflecting the more discursive and evaluative patterns of argument in the soft knowledge fields. Here persuasion is more explicitly interpretative and knowledge is typically constructed as plausible reasoning rather than as nature speaking directly through experimental findings. The presentation of research is therefore more discursive, and text-oriented bundles are heavily used to provide familiar and shorthand ways of engaging with a literature, providing warrants, connecting ideas, directing readers around the text, and specifying limitations (Hyland, 2004). About half of the text-oriented bundles in the social science texts were used to frame arguments by highlighting connections, specifying cases and pointing to limitations:

(5) *In the case of* staged financing, the problem involves double moral hazard in that the EN is inclined to shirk and the va may terminate projects too early … (BS RA)

> Most institutional talk, as will be explored later, is goal-oriented *in the sense that* the participants' behaviour is highly contingent upon their relevant identities in an institution … (AL RA)

The next most frequent group of text-oriented bundles were structuring signals, mainly used to help organise the text by providing a frame within which new arguments can be both anchored, announcing discourse goals and referring to text stages:

> (6) It is *the purpose of this chapter* to highlight some important aspects of post-allocation trading and contrast them with the conventional viewpoint. (Bus MA)
>
> *In this section we* offer evidence on the effect of corporate investment decisions on the market value of the firm. (Bus MA)

These bundles help scaffold and present arguments by considering the discoursal expectations and processing needs of a disciplinary audience.

Finally, participant bundles convey two main kinds of meaning: stance and engagement, referring to writer- and reader-focused features of the discourse respectively (Hyland, 2005). While *stance* concerns the ways writers convey epistemic and affective judgements, evaluations and degrees of commitment to what they say, *engagement* refers to writers' efforts to actively address readers as participants in the unfolding discourse.

Two thirds of all participant-oriented bundles indicated the writer's stance, and the vast majority of these were in the social science texts where personal interpretations play a far greater part in creating a convincing discourse. Most examples, in fact, express the reluctance of writers to express complete commitment to a proposition, hedging information to present it as an opinion rather than fact:

> (7) It *may be due to* the fact that vocabulary teaching has never received serious attention as one of the major concerns … (AL PhD)

> … but *it is possible that* less likely outcomes (in terms of prior probability) could have a different effect on post-choice valuation … (BS RA)

We also find these bundles expressing caution impersonally, largely through modals, epistemic adverbs and anticipatory-it patterns.

While stance bundles occurred mainly in the social science corpora, engagement bundles are largely found in hard sciences papers. These were almost all directives (Hyland, 2002), bundles which instruct readers to perform an action or to see things in a way determined by the writer. Here the writer pulls readers into the discourse to guide them to particular interpretations, typically by the use of a modal of obligation or a predicative adjective expressing the writer's judgement of necessity/importance:

(8) We conclude that, in studies on freezing-induced embolism among chaparral shrubs, *it is important to* consider the hydration of the plant …
(Bio RA)

… but *it should be noted that* in a process allowing both P and N devices to be fabricated in a well… (EE RA)

So these bundles act to position readers, requiring them to notice something in the text and thereby leading them to a particular interpretation. Their substantial presence in the hard science texts partly reflects a desire to ensure the accurate understanding of procedures and results. It also, however, represents a reluctance to adopt a more intrusive personal voice through stance options, a rhetorical choice which reduces the writer's role as interpreter and allows research to be presented as independent of any particular scientist.

VII. CONCLUSIONS AND IMPLICATIONS

In this paper we have presented corpus evidence for disciplinary variation in academic lexis, pointing to the limitations of the AWL as a general academic resource and offering a picture of academic reading and writing which emphasises the importance of disciplinary specific 4-word bundles. The different distributions of the frequency of forms and functions across disciplines helps, we believe to show something of the ways that disciplines draw on different resources to develop their arguments, establish their credibility and persuade their readers.

These findings have clear implications for EAP practitioners. Not only do they reinforce the calls by Nattinger and DeCarrico (1992), Willis (2003) and others for an increased pedagogical focus on bundles, but they also help undermine the widely held assumption that there is a single core vocabulary needed for academic study. Both individual lexical items and bundles occur and behave in dissimilar ways in different disciplinary environments and it is important that EAP materials writers and teachers recognise this, with the most appropriate starting point for instruction being the student's specific target context. In other words, we agree completely with the pedagogical principles that lay behind the AWL: that teachers should seek to teach the most relevant and useful vocabulary to their students and that corpus analyses are the best way of ascertaining this (Coxhead, 2002). Where we diverge, however, is on the nature of this vocabulary.

Numerous studies now show the extent to which language features are specific to particular disciplines, and that the best way to prepare students for their studies is not to search for universally appropriate teaching items, but to provide them with an understanding of the features of the discourses they will encounter in their particular courses. Acquisition clearly needs to be part of a well-planned and sequenced program, with a mix of explicit teaching and incidental learning, a range of activities which focus on elaboration and consolidation, and sufficient contextual and definitional information. This means, for example, encouraging learners to *notice* these items and multi-word units through repeated exposure and through activities such as matching and item identification. Consciousness raising tasks which offer opportunities to retrieve, use and manipulate items can be productive, as can activities which require learners to produce the items in their extended writing.

In sum, because academic knowledge is embedded in processes of argument and consensus-making it will always be particular to specific disciplines and their agreed ways of discussing problems. The fact that writing actually helps to *create* disciplines, rather than being just another aspect of what goes on in them, is a serious challenge to identifying overarching uniformities and encourages us to focus on what is specific in the texts our students will need to read.

REFERENCES

Biber, D. (2006). *University language: a corpus-based study of spoken and written registers*. Amsterdam: Benjamins.

Biber, D., Conrad, S. & Cortes, V. (2004). *If you look at...*: Lexical bundles in university teaching and textbooks. *Applied linguistics*. 25L, 371-405.

Carter, R. (1998). *Vocabulary: applied linguistics perspectives.* London: Routledge.

Coxhead, A. (2000) A New Academic Word List. *TESOL Quarterly*, 34:2, 213-238.

Coxhead, A. (2002). The academic word list; a corpus-based word list for academic purposes. In Ketterman, B. & Marks, G. (eds.) *Teaching and Language Corpora (TALC) conference proceedings*. Atlanta: Rodopi: pp.73-89.

Coxhead, A. & Nation, I.S.P. (2001). The specialized vocabulary of English for academic purposes. In J. Flowerdew & M. Peacock (eds.), *Research perspectives on English for Academic Purposes.* Cambridge: Cambridge University Press: pp. 252-267.

Evans, S. & Green, C. (2007) Why EAP is necessary: A survey of Hong Kong tertiary students. Journal of English for Academic Purposes. 6:1, 1-17.

Farrell, P. (1990) Vocabulary in ESP: a lexical analysis of the English of electronics and a study of semi-technical vocabulary CLCS Occasional Paper No. 25 Trinity College.

Flowerdew, J. (1993). Concordancing as a tool in course design. *System,* 21: 2, 231-244.

Halliday, MAK (1998). Things and relations: regrammaticising experience as technical knowledge. In J. Martin & R. Veel (Eds.) *Reading science*. London: Routledge, pp.185-235.

Hoey, M. (2005). *Lexical priming: a new theory of words and language*. London: Routledge.

Hyland, K. (1997). Is EAP necessary? A survey of Hong Kong undergraduates. *Asian Journal of English Language Teaching* 7, 77-99.

Hyland, K. (2002). Directives: argument and engagement in academic writing. *Applied Linguistics.* 23:2, 215-239.

Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing*. Ann Arbor: University of Michigan Press.

Hyland, K. (2005). Stance and engagement: a model of interaction in academic discourse. *Discourse Studies.* 7:2, 173-191.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes.* 27:1, 4-21.

Hyland, K. & Tse, P. (2007). Is there an 'academic Vocabulary'? *TESOL Quarterly*. 41:2.

Jordan, B. (1997). *English for Academic Purposes*. Cambridge: CUP.

Li, S.L. & Pemberton, R. (1994). An investigation of students' knowledge of academic and subtechnical vocabulary. *Proceedings joint seminar on corpus linguistics and*

lexicology, June, 1993. Hong Kong University of Science and Technology, pp. 183-196.

Nation, I.S.P. (1990) *Teaching and Learning Vocabulary* Newbury House, Mass.

Nation, I.S.P. (2001). *Learning Vocabulary in Another Language.* New York: Cambridge University Press.

Nation, I.S.P. (2002). *RANGE* (computer program). Available at http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

Nattinger, J. & DeCarrico, J. (1992). *Lexical phrases and language teaching*. Oxford: OUP.

Oakey, D. (2003). Academic vocabulary in academic discourse: the phraseological behaviour of EVALUATION in the discourse of Economics. Paper presented at *Evaluation in Academic Discourse*. University of Siena. June 16-18th 2003.

Schmitt, D. & Schmitt, N. (2005). *Focus on Vocabulary : Mastering the Academic Word List*. London: Longman.

Scott. M. & Tribble, C. (2006). *Textual patterns.* Amsterdam: Benjamins.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: OUP.

Wang, K M-T & Nation, P. (2004) Word meaning in academic English: homography in the academic word list. *Applied Linguistics*. 25:3, 291-314.

West, M. (1953). *A general service list of English words.* London: Longman.

Willis, D. (2003). *Rules, patterns and words: grammar and lexis in English language te*aching. Cambridge: CUP.

Worthington, D. & Nation, I.S.P. (1996). Using texts to sequence the introduction of new vocabulary in an EAP course. *RELC Journal* 27, 1-11.

Wray, A. & Perkins, M. (2000). The functions of formulaic language. *Language and communication*. 20, 1-28.

Yorio, C. (1989). Idiomaticity as an indicator of second language proficiency. In K. Hyltenstam & K. Obler (Eds.): *Bilingualism across the lifespan.* Cambridge: Cambridge University Press. 55-72.

