As can be seen: Lexical bundles and disciplinary variation

Ken Hyland *

Centre for Academic and Professional Literacies, Institute of Education, University of London, 20 Bedford Way, London WC1H 0AL, United Kingdom

Abstract

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 in specific contexts and contributing to our sense of coherence in a text. Bundles have begun to attract considerable attention in corpus studies in EAP, although the extent to which they differ by discipline remains an open question. This paper explores the forms, structures and functions of 4-word bundles in a 3.5 million word corpus of research articles, doctoral dissertations and Master’s theses in four disciplines to learn something of disciplinary variations in their frequencies and preferred uses. The analysis shows that bundles are not only central to the creation of academic discourse, but that they offer an important means of differentiating written texts by discipline.

1. Introduction

Multi-word expressions are an important component of fluent linguistic production and a key factor in successful language learning. While students might struggle to master phrasal verbs such as look after and sell out or idioms like in a nutshell and beat about the bush, these are relatively rare compared with the frequently occurring word sequences 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 contributing to our sense of distinctiveness in a register. Thus the presence of extended collocations like as a result of, it should be noted that, and as can be seen help identify a text as belonging to an academic register while with regard to, in pursuance of, and in accordance with are likely to mark out a legal text.

These bundles are familiar to writers and readers who regularly participate in a particular discourse, their very ‘naturalness’ signalling competent participation in a given community. Conversely, the absence of such clusters might reveal the lack of fluency of a novice or newcomer to that community. Haswell (1991, p. 236), for example, suggests that 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.

In other words, gaining control of a new language or register 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 are advantages to identifying these clusters to better help learners acquire the specific rhetorical practices of their communities.

Yet while studies point to the considerable variation of bundles in different genres (e.g. Biber, 2006; Biber, Conrad, & Cortes, 2004; Hyland, forthcoming; Scott & Tribble, 2006), how far they differ by discipline remains uncertain. This is the issue I address in this paper, examining a 3.5 million word corpus to identify the forms and functions of 4-word bundles across four contrasting disciplines.

2. Bundles, collocations and communities

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 that ‘you shall judge a word by the company it keeps’. More recently, Nattinger and DeCarrico (1992) have emphasised the importance of frequent multi-word combinations as a way of assisting communication by making language more predictable to the hearer. 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. The extensive use of such pre-fabricated sequences as it has been noted that in academic written genres, for instance, helps to signal the text register to readers and reduce processing time by using familiar patterns to link elements of new information.

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. Thus, as can be seen is a frequent and unremarkable collocation in academic writing while the equally possible as you can see or as can be observed are rarely encountered. What I shall call ‘bundles’, or frequently recurrent strings of uninterrupted word-forms, thus appear to represent a psychological association between words and reflect a very real part of users’ communicative experiences. 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, 1991:6). This extension of formulaic phrases to regular collocations such as *Have a nice day* and *I want to make three points* therefore hints at the extent of formulaicity in language use, with Altenberg (1998) suggesting that as much as 80% of natural language could be patterned in this way.

This pervasiveness 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 seeing lexical choices as constrained by the slots which grammar make available for them, they regard lexis as systematically structured through repeated patterns of use. 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. This is totally obscured by the procedures of conventional grammar.

In other words, grammar is the output of repeated collocational groupings. Sentences are typically made up of interlocking bundles as words are mentally ‘primed’ for use with other words through our experience of them in frequent associations (Hoey, 2005). Everything we know about a word is a result of our encounters with it, so that when we formulate what we want to say, the wordings we choose are shaped by the way we regularly encounter them in similar texts.

A range of corpus studies have shown how ubiquitous these bundles are in academic genres. Defining lexical bundles as combinations that recur at least 10 times per million words and across five or more texts, Biber et al. (1999, p. 994) suggest that 3-word bundles occur over 60,000 times and 4-word bundles over 5000 times per million words in academic prose. While the majority of words in any text do not occur in recurrent combinations, about 21% of the 5.3 million words of the academic component of the Longman Spoken and Written English corpus make up these common bundles, with the most frequent strings featuring over 200 times per million words.

To illustrate some of the features of these forms, Table 1 shows the most frequent 3-, 4- and 5-word bundles in my 3.5 million word corpus of academic writing in articles, PhD dissertations, and Master’s theses. The lists highlight the fact that many of the most frequent bundles in academic writing are extremely common indeed, and that these frequencies drop dramatically as strings are extended to five words and beyond. It is also clear that many 3-word bundles such as *on the other* and *it can be* frequently expand into the 5-word bundles *on the other hand the* and *it can be seen that*, supporting Cortes (2004) observation that many four and five word strings ‘hold 3-word bundles in their structure’.

Additionally, the table indicates that most bundles, unlike idiomatic phrases, are semantically transparent and formally regular, providing the building blocks of coherent discourse. They are, in other words, identified empirically purely on the basis of their frequency rather than their structure, as they typically span structural units. In particular, we might note the considerable use of what Biber et al. (1999, p. 995) call noun phrase + post modifier fragments (*the number of, the relationship between the, one of the most important*), preposition + of phrase fragments (*in terms of, on the basis of, at the beginning of the*), as well as anticipatory it fragments (*it can be, it was found that, it should be noted that*) (Hyland & Tse, 2005).
An important feature of bundles however is their variability across different genres. Biber (2006), for instance, discovered that the spoken genre of classroom teaching uses about twice as many different bundles as conversations and about four times as many as textbooks. He suggests that this extremely high density can be explained by the fact that teaching draws heavily on both oral and written genres. He also found that the bundles are required to do very different jobs in the two genres, with classroom talk comprising much higher proportions of discourse organisers (going to talk about, it has to do with) and stance bundles (I don’t know if, I want you to) than textbooks. Similarly, Cortes (2004), Scott and Tribble (2006) and Hyland (forthcoming), the latter using the corpus discussed here, also found systematic differences between genres, with bundles typical of published academic prose being far less common in writing by second language students.

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). Control of a language involves a sensitivity to the preferences of expert users for certain sequences of words over others, but students can have enormous difficulty distinguishing the idiomatic from the merely grammatical. But while there seem to be potentially enormous benefits in identifying the most frequent forms for teaching, we need to be cautious in making assumptions about the generality of academic bundles. Both Sinclair and Hoey point out, for instance, that because we all have different textual experiences, we all have a different mental concordance to draw on so that particular patterns are cumulatively loaded with the contexts we participate in. So just as

<table>
<thead>
<tr>
<th>3-Word</th>
<th>Freq.</th>
<th>4-Word</th>
<th>Freq.</th>
<th>5-Word</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>in order to</td>
<td>1629</td>
<td>on the other hand</td>
<td>726</td>
<td>on the other hand the</td>
<td>153</td>
</tr>
<tr>
<td>in terms of</td>
<td>1203</td>
<td>at the same time</td>
<td>337</td>
<td>at the end of the</td>
<td>138</td>
</tr>
<tr>
<td>one of the</td>
<td>1092</td>
<td>in the case of</td>
<td>334</td>
<td>it should be noted that</td>
<td>109</td>
</tr>
<tr>
<td>the use of</td>
<td>1081</td>
<td>the end of the</td>
<td>258</td>
<td>it can be seen that</td>
<td>102</td>
</tr>
<tr>
<td>as well as</td>
<td>1044</td>
<td>as well as the</td>
<td>253</td>
<td>due to the fact that</td>
<td>99</td>
</tr>
<tr>
<td>the number of</td>
<td>992</td>
<td>at the end of</td>
<td>252</td>
<td>at the beginning of the</td>
<td>98</td>
</tr>
<tr>
<td>due to the</td>
<td>886</td>
<td>in terms of the</td>
<td>251</td>
<td>may be due to the</td>
<td>64</td>
</tr>
<tr>
<td>on the other</td>
<td>810</td>
<td>on the basis of</td>
<td>247</td>
<td>it was found that the</td>
<td>57</td>
</tr>
<tr>
<td>based on the</td>
<td>801</td>
<td>in the present study</td>
<td>225</td>
<td>to the fact that the</td>
<td>52</td>
</tr>
<tr>
<td>the other hand</td>
<td>730</td>
<td>is one of the</td>
<td>209</td>
<td>there are a number of</td>
<td>51</td>
</tr>
<tr>
<td>in this study</td>
<td>712</td>
<td>in the form of</td>
<td>191</td>
<td>in the case of the</td>
<td>50</td>
</tr>
<tr>
<td>a number of</td>
<td>690</td>
<td>the nature of the</td>
<td>191</td>
<td>as a result of the</td>
<td>48</td>
</tr>
<tr>
<td>the fact that</td>
<td>630</td>
<td>the results of the</td>
<td>189</td>
<td>at the same time the</td>
<td>41</td>
</tr>
<tr>
<td>most of the</td>
<td>605</td>
<td>the fact that the</td>
<td>177</td>
<td>is one of the most</td>
<td>37</td>
</tr>
<tr>
<td>there is a</td>
<td>575</td>
<td>as a result of</td>
<td>175</td>
<td>it is possible that the</td>
<td>36</td>
</tr>
<tr>
<td>according to the</td>
<td>562</td>
<td>in relation to the</td>
<td>163</td>
<td>one of the most important</td>
<td>36</td>
</tr>
<tr>
<td>the present study</td>
<td>549</td>
<td>at the beginning of</td>
<td>158</td>
<td>play an important role in</td>
<td>36</td>
</tr>
<tr>
<td>part of the</td>
<td>514</td>
<td>with respect to the</td>
<td>156</td>
<td>can be seen as a</td>
<td>35</td>
</tr>
<tr>
<td>the end of</td>
<td>501</td>
<td>the other hand the</td>
<td>154</td>
<td>the results of this study</td>
<td>35</td>
</tr>
<tr>
<td>the relationship between</td>
<td>487</td>
<td>the relationship between the</td>
<td>152</td>
<td>from the point of view</td>
<td>34</td>
</tr>
<tr>
<td>in the following</td>
<td>478</td>
<td>in the context of</td>
<td>150</td>
<td>the point of view of</td>
<td>34</td>
</tr>
<tr>
<td>the role of</td>
<td>478</td>
<td>can be used to</td>
<td>148</td>
<td>it can be observed that</td>
<td>33</td>
</tr>
<tr>
<td>some of the</td>
<td>474</td>
<td>to the fact that</td>
<td>143</td>
<td>this may be due to</td>
<td>32</td>
</tr>
<tr>
<td>as a result</td>
<td>472</td>
<td>as shown in figure</td>
<td>136</td>
<td>an important role in the</td>
<td>31</td>
</tr>
<tr>
<td>it can be</td>
<td>468</td>
<td>it was found that</td>
<td>133</td>
<td>in the form of a</td>
<td>31</td>
</tr>
</tbody>
</table>
individual lexical items occur and behave in different ways across disciplines (Hyland & Tse, 2007), we need to be sure we are assisting learners towards an appropriate disciplinary-sensitive repertoire of bundles.

Applied linguists and language teachers have therefore increasingly come to see bundles as important building blocks of coherent discourse and characteristic features of language use in particular settings. But despite their importance to language production, questions remain concerning their disciplinary specific use. Analysis of specialist corpora can therefore help us to understand the kinds of language data which particular communities of users might encounter and which will inform their use. I turn to this issue now, examining bundles in the principal research genres of four contrasting disciplines.

3. Corpora and methods

Data for the study consist of three electronic corpora of written texts comprising research articles, PhD dissertations and MA/MSc theses from four disciplines (Table 2). The disciplines were chosen to represent a cross-section of academic practice: electrical engineering (EE) and microbiology (Bio) from the applied and pure sciences, and business studies (BS) and applied linguistics (AL) from the social sciences. The research article (RA) corpus consists of 120 published papers comprising 30 in the leading journals of each of the four disciplines. The PhD and Master’s corpora were written by Cantonese L1 speakers studying at five Hong Kong universities and contained 20 texts in each discipline.

I decided to focus on 4-word bundles because they are far more common than 5-word strings and offer a clearer range of structures and functions than 3-word bundles. Bundles are essentially extended collocations defined by their frequency of occurrence and breadth of use, but the actual frequency cut offs are somewhat arbitrary. This study takes a conservative approach by setting a minimum frequency of 20 times per million words and an occurrence in at least 10% of texts. Many of the higher frequency items, of course, figure far more often than this. WordSmith Tools 4 (Scott, 1996) was used to generate 4-word bundle lists for the texts in each discipline, then to concordance examples to determine their functions. I then compared the frequencies and patterns across the disciplinary corpora.

The bundles were categorised both structurally, in terms of their grammatical types, and functionally, according to their meanings in the texts. While few bundles represent complete structural units in academic writing, it is possible to group them and Biber et al’s classification was used for this purpose (Table 3).

It is also possible to identify general meanings and purposes of bundles (e.g. Biber, 2006; Cortes, 2004). Here I modified these earlier taxonomies to group bundles inductively,

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Articles</th>
<th>Doctoral</th>
<th>Masters</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electrical engineering</td>
<td>107,700</td>
<td>334,800</td>
<td>190,000</td>
<td>632,500</td>
</tr>
<tr>
<td>Biology</td>
<td>143,500</td>
<td>458,000</td>
<td>192,600</td>
<td>794,100</td>
</tr>
<tr>
<td>Business studies</td>
<td>214,900</td>
<td>437,200</td>
<td>192,300</td>
<td>844,400</td>
</tr>
<tr>
<td>Applied linguistics</td>
<td>211,400</td>
<td>670,000</td>
<td>248,000</td>
<td>1,129,400</td>
</tr>
<tr>
<td>Totals</td>
<td>677,500</td>
<td>1,900,000</td>
<td>822,900</td>
<td>3,400,400</td>
</tr>
</tbody>
</table>
in ways which seemed to best represent their functions in my corpus. The framework, discussed further below, is organised around three broad functional categories (research-oriented, text-oriented, and participant oriented), with sub-categories grouping more specific roles. Now I turn to explore these general observations in more detail by comparing the preferences of the different groups.

4. Frequencies and structures of disciplinary bundles

There were 240 different 4-word bundles altogether in the 3.5 million word corpus, totalling nearly 16,000 individual cases or just over 2% of the total words. Table 1 above shows that on the other hand was by far the most frequent of these, that it occurred 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. The top ten bundles all occurred over 60 times per million words and the analysis suggests that most bundles in academic writing are parts of noun or prepositional phrases. There are, however, some interesting disciplinary differences. The electrical engineering texts contained the greatest range of bundles with 213 different 4-word strings meeting the 20 per million words threshold (across 10% of texts), and also the highest proportion of words in the texts occurring in 4-word bundles. Biology, on the other hand, had the smallest range of bundles, the fewest examples, and the lowest proportion of texts comprised of words in bundles. Table 4 summarises this frequency information.

Many bundles used by engineers are not found in the other disciplines and there is considerably greater reliance on pre-fabricated structures than in the other fields. It is difficult to say why this might be, but speculatively it could be a consequence of the relatively abstract and graphical nature of technical communication. The density of bundles in this corpus perhaps reflects the dependence of Engineering rhetoric on visual and numeric rep-

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Most common patterns of 4-word bundles in academic writing (Biber et al., 1999, pp. 997–1025)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
<td>Examples</td>
</tr>
<tr>
<td>Noun phrase + of</td>
<td>the end of the, the nature of the, the beginning of the, a large number of</td>
</tr>
<tr>
<td>Other noun phrases</td>
<td>the fact that the, one of the most, the extent to which</td>
</tr>
<tr>
<td>Prepositional phrase + of</td>
<td>at the end of, as a result of, on the basis of, in the context of</td>
</tr>
<tr>
<td>Other prepositional phrases</td>
<td>on the other hand, at the same time, in the present study, with respect to the</td>
</tr>
<tr>
<td>Passive + prep phrase fragment</td>
<td>is shown in figure, is based on the, is defined as the, can be found in</td>
</tr>
<tr>
<td>Anticipatory it + verb/adj</td>
<td>it is important to, it is possible that, it was found that, it should be noted</td>
</tr>
<tr>
<td>Be + noun/adjectival phrase</td>
<td>is the same as, is a matter of, is due to the, be the result of</td>
</tr>
<tr>
<td>Others</td>
<td>as shown in figure, should be noted that, is likely to be, as well as the</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4</th>
<th>Bundle frequency information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discipline</td>
<td>Different bundles</td>
</tr>
<tr>
<td>Electrical engineering</td>
<td>213</td>
</tr>
<tr>
<td>Business studies</td>
<td>144</td>
</tr>
<tr>
<td>Applied linguistics</td>
<td>141</td>
</tr>
<tr>
<td>Biology</td>
<td>131</td>
</tr>
</tbody>
</table>
presentation, so while arguments are based on plausible interpretations of data, they ultimately rest on findings which are often presented in visual form. The job of language is to fashion interpretations of course, but in technical subjects it also weaves an argument by linking data or findings in routinely patterned, almost formulaic ways.

At the other end of the table, Biology employs the smallest range of different bundles and the fewest bundles overall, although the actual proportion is similar to those in the two social science disciplines. Again, the reasons for these differences are unclear, but they are related to the distinctive ways that Biology pursues and argues problems. Although Biology is like electrical engineering in that it employs visuals to buttress its arguments, this is an altogether more discursive and descriptive discipline, with a less active and applied agenda. It is also a discipline more concerned with naming and coding than the other fields, with a more specialised readership, speaking to a relatively narrow group of scientist end users with specific interests in findings which inform their own research.

In addition to different frequencies, the corpora show that the principal structures of bundles also differ across fields. Table 5 gives the percentages of the main structures in each discipline in the corpus using the Biber et al. (1999, pp. 1014-1024) classification.

As can be seen, the noun phrase with of-phrase fragment is the most common structure overall, comprising about a quarter of all forms in the corpus. This covers a range of meanings in academic discourse and in particular is widely used to identify quantity, place or size (the temperature of the, the base of the), to mark existence (a wide range of, the presence of the), or highlight qualities (the nature of the, a function of the). More interesting is the difference between disciplines, with the social science corpora making far greater use of bundles beginning with a prepositional phrase. The majority of these have an embedded of-phrase, typically indicating logical relations between propositional elements:

(1) We generated multi-item scales on the basis of previous measures, a review of the relevant literature, and interviews with marketing and purchasing personnel. (BS)

...such transformations should be studied in terms of the semantic and ideological transformations they entail. (AL)

Alternatively, in the case of up-front financing, the VC is required to provide the amount of k in a lump-sum way up front. (BS)

<table>
<thead>
<tr>
<th>Structure</th>
<th>Biology</th>
<th>Electrical engineering</th>
<th>Applied linguistics</th>
<th>Business studies</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun phrase + of</td>
<td>23.7</td>
<td>22.3</td>
<td>22.9</td>
<td>28.5</td>
<td>24.4</td>
</tr>
<tr>
<td>Passive + prepositional phrase</td>
<td>31.3</td>
<td>29.8</td>
<td>6.9</td>
<td>9.0</td>
<td>19.3</td>
</tr>
<tr>
<td>Other prepositional phrase</td>
<td>13.7</td>
<td>11.6</td>
<td>24.4</td>
<td>19.7</td>
<td>17.5</td>
</tr>
<tr>
<td>Prepositional phrase + of</td>
<td>9.2</td>
<td>7.9</td>
<td>19.9</td>
<td>16.0</td>
<td>13.5</td>
</tr>
<tr>
<td>Noun phrase + other</td>
<td>9.4</td>
<td>10.8</td>
<td>9.6</td>
<td>12.4</td>
<td>10.6</td>
</tr>
<tr>
<td>modification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>6.4</td>
<td>9.2</td>
<td>10.7</td>
<td>9.9</td>
<td>9.5</td>
</tr>
<tr>
<td>Anticipatory it structure</td>
<td>6.3</td>
<td>8.4</td>
<td>5.6</td>
<td>4.5</td>
<td>2.5</td>
</tr>
<tr>
<td>Totals</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
Here then we see the emphasis of the soft knowledge fields on the discursive exploration of possibilities and limiting conditions, identifying and elaborating relationships in argument.

The Science and Engineering texts, on the other hand, employed significantly more passive bundles, normally followed by a prepositional phrase fragment typically marking a locative or logical relation. Generally, writers here either seek to guide readers through the text (2) or to identify the basis for an assertion in an argument (3):

(2) The experiment setup is shown in Fig. 4.13. (EE)
   The value of Rs is given by Eq. (3.11). (EE)
   All important events for pot trials are summarised in Table 4.11. (Bio)

(3) This apparent stability might be due to the complexing of plasma/serum DNA with proteins in the circulation. (Bio)
   The measurement is based on the evaluation of infrared images produced by thermal waves. (EE)
   Therefore, an antisense approach can be used to block the expression of endogenous agrin in PC12 cells. (Bio)

Identifying tabular or graphic displays of data and the bases of an assertion are typically constructed through formulaic passive constructions in the hard sciences. This both highlights the research or text feature being discussed and can help downplay the personal role of the scientist in the interpretation of data to suggest that the results would be the same whoever conducted the research.

Interestingly, the science writers also tended to employ more examples of the anticipatory-it pattern, which is another means of disguising authorial interpretations. These bundles introduce extraposed structures and function to foreground the writer’s evaluation without explicitly identifying its source:

(4) It is possible that an increase in ethylene production in these fruits is mediated by CABA. (Bio)
   It is found that the optimal number of processing elements is application-dependent. (Bio)
   Referring to Fig. 1, it can be seen that each stage of the early bucket brigade circuit consists of an NPN bipolar transistor and a storage capacitor. (EE)

I now turn to look at the patterns themselves and their distributions across the disciplines.

5. Patterns and variations

There were also considerable differences in the 4-word bundles themselves across disciplines. Table 6 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 italicized.

The table may make depressing reading for commercial materials writers seeking to identify universals of academic writing and compile word lists for general academic pur-
poses. 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. Applied linguistics has 29 items in the top 50 which do not occur in any of the other lists and electrical engineering has 28. The discipline-specificity of these preferences for 4-word bundles is illustrated by
the bold and italicized 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 EAP course 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.

The greatest affinity is between broadly cognate fields, as business studies and applied linguistics share 18 items, although only on the basis of, in the context of, the relationship between the, and it is important to are exclusive to these two fields. 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.

While consideration of the lexical composition of these formulaic strings is useful, we are better able to understand the roles they play in academic discourse by examining their discourse function and I turn to this in the next section.

6. Functions of bundles

A framework for analysing the bundles found in this corpus was developed from Biber’s (Biber, 2006; Biber et al., 2004) classification. While my main categories are similar, differences in the two corpora necessitated modifications. Biber’s taxonomy emerged from a much broader corpus of spoken and written registers which included casual conversation, textbooks, course packs, service encounters, institutional texts, and so on, and this seems to have yielded far more personal, referential, and directive bundles than my more research-focused genres. Biber, for instance employs stance as a super-ordinate category while I have folded it into a grouping in which bundles refer to either the writer or reader. In addition, the realisations in different categories are so different that use of the same sub-groups would invite unproductive comparisons. This classification therefore collects bundles into the three broad foci of research, text and participants, and introduces sub-categories which specifically reflect the concerns of research writing. These are:

**Research-oriented** – help writers to structure their activities and experiences of the real world includes:

- **Location** – indicating time/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, 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 includes:
• **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 figure*).

• **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) includes:

• **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*).

Using this classification scheme, we find some functional categories are strongly connected to the structural patterns discussed earlier, with noun phrases + of structures prominent in research-oriented functions, prepositional phrase patterns in text-oriented functions, and *anticipatory it* largely occurring in participant functions. We can also note a roughly even split between research- and text-oriented bundles overall, with participant strings being far less frequent. Table 7, however, shows that once again there are differences in disciplinary distributions, pointing to variations in what writers are attempting to achieve through their linguistic choices.

### 6.1. Research-oriented bundles

One clear difference is the greater concentration 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 scale of this use functions to impart a greater real-world, laboratory-focused sense to writing in the hard sciences. Many of these bundles contributed to the description of research objects or contexts, specifying aspects of models, equipment, materials or aspects of the research environment, and were typically realised by noun phrase + of structures:

(5) *The structure of the* coasting-point identification model (see Fig. 5.6) *can be divided into the following areas for description.*  

**Table 7**

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Research-oriented</th>
<th>Text-oriented</th>
<th>Participant-oriented</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology</td>
<td>48.1</td>
<td>43.5</td>
<td>8.4</td>
<td>100</td>
</tr>
<tr>
<td>Electrical engineering</td>
<td>49.4</td>
<td>40.4</td>
<td>9.2</td>
<td>100</td>
</tr>
<tr>
<td>Applied linguistics</td>
<td>31.2</td>
<td>49.5</td>
<td>18.6</td>
<td>100</td>
</tr>
<tr>
<td>Business studies</td>
<td>36.0</td>
<td>48.4</td>
<td>16.6</td>
<td>100</td>
</tr>
<tr>
<td>Overall</td>
<td>41.2</td>
<td>45.5</td>
<td>13.2</td>
<td>100</td>
</tr>
</tbody>
</table>
...the performance of the coder is less affected by neither improper voiced-unvoiced classification nor voiced-unvoiced speech transitions of different durations. (EE)

The size of the perforations becomes progressively smaller towards the base of the apparatus. (Bio)

Over half of all cases, however, were used to depict research procedures, showing the ways that experiments and research were conducted:

(6) The DNA was precipitated in the presence of 2.5 volumes of ethanol and 0.1 volume of 3.0 M sodium acetate pH. (Bio)

Transmission phase angle modulation can be used to increase the stability of the system, by maintaining the angle at a low value. (EE)

All of the precipitate was added to the cells in a 100 mm culture plate or 300 mm of the precipitate to a 60 mm culture plate. (Bio)

This emphasis on the ways the research was conducted plays an important role in conveying the grounded, experimental basis of research in the hard sciences. The physical practicalities of scientific study played a far greater part in the student discourses than in the articles, however, perhaps reflecting the ways that the Master’s level students 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 demonstrate their ability to handle research methods appropriately and stake a claim to being comfortable with the subject content of the discipline. Consequently, bundles which set out procedures comprise a far higher proportion in the theses and dissertations, as do those which refer to the specific topic or context of the research:

(7) Thus by studying this type of faults, the transient stability of the power system under the most adverse condition can be determined. (EE MSc)

This can improve the signal-to-noise ratio of the reconstruction and the reconstructed signal will become more natural. (EE MSc)

Forty percent of the total land area in Hong Kong is designated as country parks which together covers an area of over 40,000 ha. (Bio MSc)

These patterns may, therefore, reveal the preoccupations of the apprentice, and perhaps specifically the second language apprentice, demonstrating competence through the control of physical resources and disciplinary research practices. But the significantly greater use of research-oriented bundles in the hard knowledge fields also expresses something of a scientific ideology which emphasises the empirical over the interpretive, minimising the presence of researchers and contributing to the “strong” claims of the sciences. Highlighting research rather than its presentation places greater burden on research practices and the methods, procedures and equipment used, and this allows scientists to emphasise demonstrable generalisations rather than interpreting individuals. New knowledge, then, is accepted on the basis of empirical demonstration and experimental results designed to test hypotheses related to gaps in knowledge. The rhetorical conventions of the field, including the preferred patterns of 4-word bundles, help contribute to this epistemological framework.
6.2. Text-oriented bundles

In contrast, the applied linguistics and business studies corpora were dominated by text-oriented strings, which were particularly marked in the research articles where they comprised almost two-thirds of all bundles. This reflects the more discursive and evaluative patterns of argument in the soft knowledge fields, where persuasion is more explicitly interpretative and less empiricist, producing discourses which often recast knowledge as sympathetic understanding, promoting tolerance in readers through an ethical rather than cognitive progression (Hyland, 2004). So while claims are often based on observations of real-world phenomena, knowledge is typically constructed as plausible reasoning rather than as nature speaking directly through experimental findings. The presentation of research is therefore altogether more discursively elaborate, 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.

Perhaps not surprisingly, about 50% of text-oriented bundles in the social science texts worked to frame arguments by highlighting connections, specifying cases and pointing to limitations:

(8) The term ‘linguistics’ might be too narrow *in terms of the* diverse knowledge-base and expertise that is required in the applied linguist’s job. (AL)

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)

The levels are connected *in the sense that* it is impossible to appreciate the functioning at any one level without taking account of the other levels. (AL)

These bundles tend to be preposition + of structures and are used to focus readers on a particular instance or to specify the conditions under which a statement can be accepted, working to elaborate, compare and emphasise aspects of an argument.

While framing devices also comprised a high proportion of text-oriented bundles in the hard science corpora, there were far more in the applied linguistics and business texts. Here readers are often drawn from a wider knowledge base and include both those from other specialisms and disciplines and practitioners looking to apply the research in different areas. This readership is not only less cohesive than in the sciences but the research often has to be contextualised far more carefully and the connections between components explained in greater detail for readers unfamiliar with the thread of prior research.

The next most frequent group of bundles in the text-oriented category were structuring signals. A substantial portion of these help organise the text by providing a frame within which new arguments can be both anchored and projected, announcing discourse goals and referring to text stages:

(9) The *purpose of this paper* is to investigate the perceptions of consumers in the Hong Kong market toward a foreign service offering, specifically fast food. (BS)

*In this chapter we* introduce a forecasting technique utilizing the notion of, global optimization to define the input-output membership functions with respect to. . . . (EE)
In this section we offer evidence on the effect of corporate investment decisions on the market value of the firm. (BS)

These bundles help frame, scaffold, and present arguments as a coherently managed and organised arrangement, reflecting writers’ awareness of the discursive conventions of a sustained discussion and in consideration of the discoursal expectations and processing needs of a particular disciplinary audience. They are especially widespread in the much longer doctoral texts, where they help to structure arguments over a greater span of text. As Bunton (1999, p. 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.

Equally, however, these bundles represent an awareness of both argument and audience, and their use suggests writers’ attempts to position themselves as competent academics able to control the rhetorical conventions of their fields.

Another group of structuring signals point to other parts of the texts to make additional material salient and available to the reader in recovering the writer’s intentions. As mentioned earlier, the electrical engineers were particularly heavy users of these signals, reflecting their dependence on graphical and numerical information and the need to refer to these in their arguments:

(10) 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)

As shown in Fig. 2, VDSATH is approximately equal to VDS when the transistor operates in the triode region. (EE)

As shown in the example, process steps can be parameterised with materials object names. (EE)

Biologists, on the other hand, made considerable use of resultative markers, bundles which introduce writer’s interpretations and understandings of research processes and outcomes. This is a key function in the rhetorical presentation of research as these bundles signal the main conclusions to be drawn from the study and highlight the inferences the writer wants readers to draw from the discussion:

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

This is due to the precipitation of solid state CdS in the anoxic paddy soil. (Bio)

These results suggest that the observed variability is largely statistical, but that spatial variations cannot be entirely neglected. (Bio)

Resultative markers can frame an assertive construal of events, boosting the writer’s position and directing readers to a categorical understanding, but more often they preceded a more conciliatory stance, as the last example here, 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. Such considerations are at the heart of participant-oriented selections.
6.3. Participant-oriented bundles

Participant bundles provide a structure for interpreting a following proposition, conveying two main kinds of meaning: stance and engagement. These labels refer to writer- and reader-focused features of the discourse respectively, representing key aspects of interaction in texts (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.

Some two thirds of all participant-oriented bundles indicated the writer’s stance, and the overwhelming majority of these were in the social science texts. Here, writers have to establish their claims through more explicit evaluation and engagement: personal credibility, and explicitly getting behind arguments, plays a far greater part in creating a convincing discourse for these writers:

(12) Such a dilemma may be due to the fact that they generally are unable to get support on English difficulties. (AL)

Ventures with superior performance are more likely to keep the original designs or even develop towards separate entities. (BS)

Nevertheless, it is possible that greater social interaction between marketing and Engineering managers would be beneficial to organizational interests. (BS)

These few examples not only illustrate the use of stance bundles in the social science texts, but also the fact that they largely convey a reluctance to express complete commitment to a proposition, allowing writers to present information as an opinion rather than accredited fact. Hedges figure prominently here as do the anticipatory-it structures discussed above. These realisations help to protect the writer from possible false interpretations and indicate the degree of confidence that it may be prudent to attribute to the accompanying statement.

Not only are these stance bundles largely used to communicate uncertainty or caution, but they are also almost entirely expressed impersonally; in fact there is only one personal stance structure in the entire corpus, found in the applied linguistics collection:

(13) In concluding this chapter, I would like to emphasize that this study does not reject any theories proposed in previous studies on code-switching. (AL)

Finally, I would like to suggest that the teaching of LSP should re-assess its current emphasis on the differences between professional groupings. (AL)

More usually, however, stance is expressed impersonally through bundles which employ models, epistemic adverbs and anticipatory-it patterns, as in example (12) above.

While stance bundles occurred principally in the social science corpora, and here overwhelmingly in the research articles, writers in the hard sciences largely employed strings which sought to engage readers. These were almost all directives (Hyland, 2002), bundles which explicitly mark the presence of the ‘reader-in-the-text’ (Thompson, 2001) and instruct readers to perform an action or to see things in a way determined by the writer. Here the writer pulls the audience into the discourse at critical points 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:

(14) Intuitively, we can see that if the income levels of two economies become more similar over time, it must be the case that the poor economy is growing faster. (BS)

It should be noted that the extracted MAPs are associated with the polymerized tubulin. (Bio)

Second, it is important to recognize that the current state of knowledge in this area is still in its infancy. (AL)

In other words, although mixtures of zero al exists, it is necessary to carefully optimize the material parameters associated with the rotational viscosity. (EE)

Here the writer acknowledges the dialogic dimension of research writing, intervening to direct the reader to some action or understanding. These bundles therefore act to position readers, requiring them to notice something in the text and thereby leading them to a particular interpretation.

The relatively substantial presence of these items in the hard science corpora reflects the fact that these disciplines place considerable emphasis on precision, particularly to ensure the accurate understanding of procedures and results. The more linear and problem-oriented approach to knowledge construction found in the sciences allows arguments to be formulated in highly standardised, almost shorthand, ways which presuppose a degree of theoretical knowledge and routine practices not possible in the soft fields. As a result, directives offer writers an economical and precise form of expression which cuts more immediately to the heart of technical arguments. This high proportion of engagement bundles, however, also represents a reluctance to adopt a more intrusive personal voice through stance options, a rhetorical choice which reduces the writer’s role as agent and interpreter and allows research to be presented as independent of any particular scientist.

I should point out that participant bundles were predominantly a feature of the research articles and that virtually all cases in the two student genres were examples of engagement. This avoidance of participant-oriented bundles may be a result of my student corpus and perhaps reflect the influence of a second language factor on these patterns. It certainly underlines a preference for impersonality by Hong Kong students found in other studies, which seems to result from both educational experiences and cultural preferences for a conciliatory, non-interventionist stance (Scollon & Scollon, 1995). While it is worth mentioning that stance and engagement are often expressed in other ways than 4-word bundles (e.g. Biber, 2006; Hyland, 2005), the relative absence of their use in the student corpus suggests that these writers may be uncomfortable in 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 one they do not wish to take.

7. Conclusions

My main purpose in this study has been to explore the extent to which phraseology contributes to academic writing by identifying the most frequent 4-word bundles in the key genres of four disciplines. The findings support studies by Cortes (2004) and Biber (2006) which show considerable variations in the frequency of forms, structures and func-
tions across types of academic writing, but extend these studies by examining several disciplines and relating variations in the social and rhetorical practices of academic communities. The study indicates that writers in different fields draw on different resources to develop their arguments, establish their credibility and persuade their readers, with less than half of the top 50 bundles in each list occurring in any other list.

The results need to be treated with some caution, of course. I have not discussed the possible influence of first language on the findings in any detail and a corpus of first language students might well suggest different preferences, although at this level of proficiency I would be surprised if this were the case. I am also aware of the limitations of the size of my sample, as 3.5 million words is not a large corpus in terms of work being conducted today. More work with different disciplines, genres, and first language groups is likely to yield a fuller picture of community-specific practices. I hope, however, that I have done enough here to suggest that 4-word bundles should be regarded as a basic linguistic construct and that their distributions can help characterise disciplinary discourses.

While there is little space remaining for elaboration, these findings have clear implications for EAP practitioners. Not only do they reinforce the calls by Nattinger and DeCarrico (1992), Lewis (1997), 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. Bundles occur and behave in dissimilar ways in different disciplinary environments and it is important that EAP course designers recognise this, with the most appropriate starting point for instruction being the student’s specific target context. Corpus-informed lists and concordances can be used to help establish frequently occurring and otherwise productive bundles for EAP courses and the design of relevant teaching materials. It is important, however, that these lists and concordances are derived from the genres students will need to write and read. This means, for example, encouraging learners to notice these 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.

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. The further study of bundles, I suggest, can offer insights into a crucial, and often overlooked, dimension of genre analysis and help provide us with a better understanding of the ways writers employ the resources of English in different academic contexts.

References


