Annual Review of Applied Linguistics (2012), **32**, 150–169.

doi: 10.1017/S0267190512000037

Bundles in Academic Discourse

Ken Hyland

Automated, frequency-driven approaches to identifying commonly used word combinations have become an important aspect of academic discourse analysis and English for academic purposes (EAP) teaching during the last 10 years. Referred to as clusters, chunks, or bundles, these sequences are certainly formulaic, but in the sense that they are simply extended collocations that appear more frequently than expected by chance, helping to shape meanings in specific contexts and contributing to our sense of coherence in a text. More recently, work has extended to "concgrams," or noncontiguous word groupings where there is lexical and positional variation. Together, these lexical patterns are pervasive in academic language use and a key component of fluent linguistic production, marking out novice and expert use in a range of genres. This article discusses the emerging research which demonstrates the importance of formulaic language in both academic speech and writing and the extent to which it varies in frequency, form, and function by mode, discipline, and genre.

An important component of fluent linguistic production is control of the multiword expressions referred to as clusters, chunks, or lexical bundles. While perhaps not strictly formulaic by Wray's (2002) definition, which makes a claim that sequences are stored in the mental lexicon, these strings are nevertheless glued together in everyday discourse. Simply put, bundles are statistically the most frequent recurring sequences of words in any collection of texts: extended collocations that appear more repeatedly than expected by chance (Biber, Johansson, Leech, Conrad, & Finegan, 1999). They are made evident through corpus analysis software that retrieves multiword units with specified frequency and distribution criteria and as a result are neither idiomatic nor, usually, complete grammatical units (Biber, 2006), throwing up strings such as *it was found that* and *in the case of*. They are familiar to users of a language and have customary pragmatic or discoursal functions. The criterion of frequency is therefore paramount and distinguishes bundles from, say, Renouf and Sinclair's (1991) *collocational frameworks* of productive preselected patterns and from fixed idioms.

While some research has been published focusing on academic bundles in languages such as Spanish (e.g., Butler, 1998; Cortes, 2008; Tracy-Ventura, Cortes,

& Biber, 2007) and Korean (Kim, 2009), published work on other languages is limited to particular language groups writing in English, such as Chinese (e.g., Ma, 2009; Wei, 2007; Xu, 2007). The vast majority of research looks at academic bundles in English, and therefore this chapter focuses on English, discussing the emerging research which demonstrates the importance of this type of formulaic language in both academic speech and writing and the extent to which it varies in frequency, form, and function by mode, discipline, and genre.

IDENTIFYING BUNDLES: FREQUENCY, DISTRIBUTION, AND VARIABILITY

Research into lexical bundles follows the pioneering work of Bengt Altenberg (1993, 1998), who created the methodology to identify frequency-defined recurrent word combinations and who combined grammatical and functional analysis in categorizing them. Clearly an approach to identifying and classifying formulaic units based solely on frequency of occurrence and breadth of use has the advantage of being methodologically clear-cut, although researchers have used different frequency and distribution criteria.

The threshold frequency, which determines the number of bundles to be included in the list, has ranged from 10 (Biber et al., 1999; Biber, 2006) to 20 (Cortes, 2004; Hyland, 2008a, 2008b) to 40 times per million words (Biber, Conrad, & Cortes, 2004). Such normalization methods, which are widely used to compare individual words across different sized corpora, may, however, be unreliable when working with lexical bundles, and more research is needed to establish their validity. Moreover, analysts using smaller spoken corpora often employ much lower cutoffs (De Cock, 1998; Nesi & Basturkmen, 2006), but it can be very problematic to determine what a bundle is in very small corpora. This raises a larger issue of using small samples in the study of bundles, as small corpora tend to produce many more bundles than their larger counterparts in the same registers. Thus, further research is needed before reliable comparisons are made.

A second identification criterion is that sequences have to occur in a specified number of files in the corpus, such as three to five texts (e.g., Biber & Barbieri, 2007) or 10 percent of texts (Hyland, 2008a) to avoid the quirks of individual speakers or writers. Finally, analysts must decide on the length of strings they select. Three-word bundles are extremely common, and tend not to be very interesting, while 5- and 6-grams are comparatively rare and often subsume shorter ones. Four-word bundles seem to be most often studied, perhaps because they are over 10 times more frequent than five-word sequences and offer a wider variety of structures and functions to analyze. Biber et al. (1999), in fact, suggested that four-word bundles and above "are more phrasal in nature and correspondingly less common" (p. 992).

In terms of analysis, researchers often manually exclude bundles with noun phrases as being too text-dependent and remove overlapping word sequences where two four-word bundles are actually part of a five-word string (e.g., *it has been suggested* and *has been suggested that*; Chen & Baker, 2010). Frequency analysis, moreover, produces long lists of recurrent word sequences that often

run counter to intuition. Sequences such as *on the other hand* and *the results suggest* appear psycholinguistically unproblematic compared to *at the end of* and *is one of the*, which have similar frequencies. Some researchers have therefore chosen to weed out nonintuitive expressions to produce shorter lists which include only units of "structural and idiomatic coherence" (Simpson, 2004, p. 42), although this is a method vulnerable to claims of subjectivity.

Others have relied upon complex combination of "corpus statistics, linguistic analyses, psycholinguistic processing metrics and instructor insights" to produce "psycholinguistically salient sequences" for teaching purposes (Simpson-Vlach & Ellis, 2010, p. 490). One aspect of this is often the mutual information (MI) score, which is a statistical measure of association between words in a bundle. Programs such as Collocate (Barlow, 2004) compute this score automatically to indicate the strength of collocations, comparing the frequency of a word combination to the overall frequencies of each of the individual words. The method has been used in several studies (e.g., Ellis, Simpson-Vlach, & Maynard, 2008; Simpson-Vlach & Ellis, 2010) as it appears to offer an indication of phrasal coherence, corresponding to distinctive functions or meanings. MI scores, however, were originally conceived for two-word collocations and may be unreliable when trying to account for the frequency of longer expressions. It tends to privilege low-frequency items and simply reflects a likelihood that a pair of words will occur together, regardless of order (Biber, 2009).

Moreover, the automated, frequency-driven means of retrieving lexical bundles allows us to say little about how such sequences are psycholinguistically processed (e.g., Wray, 2002) or acquired (e.g., Schmitt, Dornyei, Adolphs, & Durow, 2004). Their recurrence in multiple texts by different users, however, suggests at least some perceptual salience among users and conventionalization within a particular discourse community. The fact they are identified through corpus-driven research means that they emerge inductively from analysis of a corpus rather than the a priori assumptions of the analyst. Indeed, they are a key way of shaping 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.

Clearly, bundles refer only to fixed collocational patterns, yet our intuitions suggest that there is considerable positional flexibility in formulaic sequences, and Biber (2009) noted various pattern types in particular four-word combinations. While software like WordSmith Tools 5 (Scott, 2008) is able to generate high-frequency phrases such as *the relationship between the*, it misses instances of the same pattern in, for example, *the clear relationship between the* or *the uncertain relationship between the.* Clearly, some sequences have optional slots in addition to their fixed elements and these remain undiscovered. By revealing noncontiguous word groupings, or concgrams, recent software developments seek to overcome this limitation (Cheng, 2007; Cheng, Greaves, Sinclair, & Warren, 2007; Greaves, 2009).

According to the program designer, Chris Greaves, a concgram is "all of the permutations of constituency variation and positional variation generated by

the association of two or more words" (Cheng, Greaves, & Warren, 2006, p. 414). This is, however, a relatively new way of identifying and categorizing word associations that has yet to generate published studies of academic discourse. Preliminary searches of nonacademic spoken corpora have found that the majority of concgrams are composed of noncontiguous collocations, revealing both constituency (AB, ACB) and positional (AB, BA) variations. There is clearly great potential here to illuminate the formulaic patterning, especially phraseological variation, of academic speech and writing.

THE IMPORTANCE AND DISTINCTIVENESS OF ACADEMIC SEQUENCES

These sequences are important to writers and speakers for at least three reasons (Coxhead & Byrd, 2007):

(1) Their repetition offers users (and particularly students) ready-made sets of words to work with.
(2) They help define fluent use and therefore expertise and legitimate disciplinary membership.
(3) They reveal the lexico-grammatical community-authorized ways of making-meanings.

Routinely employed sequences, therefore, work to facilitate pragmatically efficient communication, and in academic discourse often function to structure a discourse by guiding readers through a text (*in the next section*, *as shown in figure*) or by linking ideas (*is due to the*, *in contrast to*). In addition, by signaling appropriate use of disciplinary resources, they allow writers to display solidarity with colleagues (Cortes, 2006) and to construct a disciplinary competent voice (Hyland, 2008a; Pang, 2010).

Lexical bundles, therefore, seem to reflect a very real part of users' communicative experiences. As suggested by Sinclair's (1991) idiom principle, there is a phraseological tendency in language use whereby speakers and writers co-select words in routine ways. 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 come across them in similar texts. Needless to say, these different kinds of lexical patterns are pervasive in academic language use and a key component of fluent linguistic production, marking out novice and expert use in both spoken and written contexts.

Corpus research has identified recurrent patterns in corpora of written and spoken language which occur significantly more frequently in academic than in other, nonacademic registers. This suggests, for example, that academic writing draws on a much larger stock of prefabricated phrases than either news or fiction in the British National Corpus Baby edition, with over 450 different four-word clusters occurring more than 10 times in one million words (Hyland, 2008a); see Table 1.

Table 1. Ranked Four-Word Bundles in the British National Corpus Baby Edition by Frequency

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

Clearly, in this corpus at least, academic writing shares only a few clusters with either fiction or conversation. These kinds of register differences are confirmed by Biber et al. (1999) and Simpson-Vlach and Ellis (2010) with much larger corpora. In seeking to identify high-frequency academic-specific bundles for teaching purposes, for example, Simpson-Vlach and Ellis list over 200 three-, four-, and five-word bundles which are statistically more common in academic texts than in a large corpus of 15 nonacademic spoken and written genres. The most statistically more frequent being *in terms of*, *at the same time*, and *from the point of view*.

Biber et al. (1999) showed that this distinctiveness extends to the formal properties of bundles, so that academic bundles are frequently preposition + noun phrase fragments, noun phrase + *of* phrase fragments (see also Hyland, 2008b; Scott & Tribble, 2006) or anticipatory *it* fragments (Hyland & Tse, 2005). Together, these three forms make up over 70 percent of four-word patterns in academic discourse but rarely figure in conversation, where 60 percent of

Table 2. Common Forms of Four-Word Bundles in Academic Writing

Structure	Examples
noun phrase + of	the end of the, the nature of the, the beginning of the, a large number of
other noun phrases	the fact that the, one of the most, the extent to which,
prepositional phrase + of	at the end of, as a result of, on the basis of, in the context of
other prepositional phrases	on the other hand, at the same time, in the present study, with respect to the
passive + prep phrase fragment	is shown in figure, is based on the, is defined as the, can be found in
anticipatory it + verb/adj	it is important to, it is possible that, it was found that, it should be noted
be + noun/adjectival phrase	is the same as, is a matter of, is due to the, be the result of
others	as shown in figure, should be noted that, is likely to be, as well as the

Note. Adapted from Biber et al. (1999, pp. 997–1025).

patterns are personal pronoun + lexical verb phrases (*I don't know what, I thought it was*) and auxiliary + active verb (*have a look at, do you want a*). These patterns are therefore strong register discriminators. Table 2 shows the most common patters in academic writing.

FORMULAIC PATTERNS IN SPOKEN AND WRITTEN ACADEMIC DISCOURSE

Corpus studies have also shown how ubiquitous these bundles are in academic genres. Defining lexical bundles as combinations that recur at least 10 times per million words across five or more texts, Biber et al. (1999) suggested that three-word bundles occur over 60, 000 times and four-word bundles over 5,000 times per million words in academic prose. The lists highlight the fact that many of the most frequent bundles in academic writing are extremely common indeed, and like bundles in other registers, that these frequencies drop dramatically when we look at strings of five words or more. *On the other hand* was by far the most frequent cluster, which occurred 100 times per million words and was more than twice as common as those next placed, *at the same time* and *in the case of*. The top 10 all occurred more than 60 times per million words, and the entire list was dominated by prepositional phrase constructions and noun phrases with *of* fragments.

The most frequent three-, four- and five-word bundles in a 3.5-million word corpus of articles, PhD dissertations, and master's theses are shown in Table 3 (Hyland, 2008b).

Table 3. Most Frequent Three-, Four-, and Five-Word Bundles in Academic Articles and Theses

3-Word	Freq.	4-Word	Freq	5-Word	Freq.
in order to	1,629	on the other hand	726	on the other hand the	153
in terms of	1,203	at the same time	337	at the end of the	138
one of the	1,092	in the case of	334	it should be noted that	109
the use of	1,081	the end of the	258	it can be seen that	102
as well as	1,044	as well as the	253	due to the fact that	99
the number of	992	at the end of	252	at the beginning of the	98
due to the	886	in terms of the	251	may be due to the	64
on the other	810	on the basis of	247	it was found that the	57
based on the	801	in the present study	225	to the fact that the	52
the other hand	730	is one of the	209	there are a number of	51
in this study	712	in the form of	191	in the case of the	50
a number of	690	the nature of the	191	as a result of the	48
the fact that	630	the results of the	189	at the same time the	41
most of the	605	the fact that the	177	is one of the most	37
there is a	575	as a result of	175	it is possible that the	36
according to the	562	in relation to the	163	one of the most important	36
the present study	549	at the beginning of	158	play an important role in	36
part of the	514	with respect to the	156	can be seen as a	35
the end of	501	the other hand the	154	the results of this study	35
the relationship between	487	the relationship between the	152	from the point of view	34
in the following	478	in the context of	150	the point of view of	34
the role of	478	can be used to	148	it can be observed that	33
some of the	474	to the fact that	143	this may be due to	32
as a result	472	as shown in figure	136	an important role in the	31
it can be	468	it was found that	133	in the form of a	31

It is also clear that many four- and five-word strings, such as *on the other hand the* and *it can be seen that* "hold three word bundles in their structure" (Cortes, 2004; p. 401), thus suggesting that three- and four-grams might offer a more productive focus for teachers and analysts. Table 3 also shows that most bundles, unlike idiomatic phrases, are semantically transparent and formally regular, many being nominal or prepositional phrases (cf. Butler, 1998). In particular, we can see the considerable use of what Biber et al. (1999) call noun phrase + postmodifier 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*).

Studies of bundles in spoken academic discourse have been much rarer and mainly limited to the work conducted by Biber and colleagues at Northern Arizona University (e.g., Biber, 2006; Biber & Barbieri, 2007; Biber et al., 2004; Cortes & Csomay, 2007). This research has investigated a range of genres (or "registers" in Biber's parlance) including both instructional (classroom teaching, study groups) and noninstructional contexts (student advising, office hours, class management, and university service encounters). This research shows that while classroom teaching uses an extremely wide variety of different bundles in comparison to conversation, textbooks, and academic prose (Biber et al., 2004), these bundles are even more prevalent and diverse in noninstructional genres such as classroom management and service encounters (Biber & Barbieri, 2007). Results such as this, however, need to be seen in the context of the preceding comments concerning the reliability of frequencies generated from very small corpora.

There is also a substantial reliance on what Biber called stance bundles, concerned with expressing epistemic evaluations, attitudes, or modal meanings and with framing new propositional information (examples from Biber, 2006):

(1) *I want you to* take out a piece of paper.
Right now *what we're going to* take a look at are ones that are [. . .] positive and beneficial.
All *you have to do* is work on it.

Cortes and Csomay (2007) suggested that these stance bundles are found particularly at the beginning of university lectures, where teachers are trying to negotiate class management issues, and toward the middle, where they are eliciting class participation. Discourse organizing bundles are also very common in classroom teaching—and in conversation—mainly to introduce and elaborate topics:

(2) What I want to do is quickly run through the exercise . . .
Today *we are going to talk* about testing hypotheses.
It *has to do with the* START talks, with the Russians.

Simpson (2004) confirmed the importance of interactive expressions in her study of the MICASE (Michigan Corpus of Academic Spoken English) corpus, but highlighted the significance of discourse organizing bundles, particularly those use to summarize, sequence, and focus information. Simpson also, however,

noted the influence of idiolect and speech event on distributions. Her data also showed a considerable variation in the expressions favored by professors (*and so on, in other words, and so forth*) and by students (*I was like, something like that, you know what I mean*) in this U.S. university context.

BUNDLES AND GENRE VARIABILITY

Despite these apparent differences between spoken and written discourse observed by Biber and colleagues, already mentioned, it is genre, rather than mode, which is more important in distinguishing the distribution of bundles. Biber and Barbieri (2007) made this clear:

> The extent to which a speaker or writer relies on lexical bundles is strongly influenced by their communicative purposes, in addition to general spoken/written differences. The explanation for the infrequent use of lexical bundles in the academic written registers (textbooks and academic prose) apparently lies in the restricted communicative goals of those registers—focused on informational communication—rather than the written mode per se. (p. 273)

An important feature of bundles is, therefore, their variation across different genres, and this, in turn, contributes to our understanding of the integrity of generic patterning.

Biber (2006), for example, shows us that the spoken genre of classroom teaching uses about twice as many different bundles as conversation and about four times as many as textbooks. Biber suggested that this extremely high density could 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 organizers (*going to talk about*, *it has to do with*) and stance bundles (*I don't know if*, *I want you to*) than textbooks. Similarly, Simpson (2004) and Simpson and Mendis (2003) discovered, perhaps unsurprisingly, almost completely different sets of bundles in monologic (lectures) and dialogic (tutorials, class discussions) genres, with more than twice as many expressions in the interactive speech events (*I'll show you*, *in a minute*, *in some sense*).

This genre variation is repeated in written genres, particularly in published academic papers and student texts. Chen and Baker (2010), for example, discovered a considerable "gap between native expert academic prose and immature student academic writing" (p. 34). This is particularly marked in the high uses of referential bundles, which are used to specify attributes of various kinds in three different ways:

- Framing: *in the context of, the existence of the*
- Quantifying: *a wide range of, the extent to which*
- Place/time/text—deictic: *are shown in figure, at the same time*

The student texts, on the other hand, contained far more discourse organizers. Chen and Baker (2010) attributed these variations to both proficiency and genre differences, noting more so-called native-like writing among the advanced learners in the corpora. Similarly, Cortes (2004) found that the bundles used by students did not correspond to those employed by professional authors, and that many bundles frequently found in published papers were never used by students at all.

Seeking to control for proficiency, I explored a corpus of 3.5 million words of skilled writing, looking at published articles and at high graded master's theses and doctoral dissertations by second language (L2) writers in Hong Kong (Hyland, 2008a, 2008b). There were considerable differences, with the articles containing 71 different four-word bundles of 20 per million words or more in more than 10 percent of texts; the PhD dissertations, 95 different clusters; and the master's texts, 149. Overall, in fact, the postgrad genres appear to be more phrasal than the published one, with four-word bundles composing 5.1 percent of the master's theses, 3.8 percent of the PhD dissertations, and 3.1 percent of the research articles. While this may suggest a certain conservatism among students and an attempt to rely on less risky prefabricated language (e.g., Hyland & Milton, 1997), it is also true that the research article has a different purpose, audience, and repertoire of rhetorical features compared to the student genres, representing what Swales (1990) referred to as a *norm developing* practice, concerned with persuasive reporting through engagement with the professional world, rather than *norm developed* which largely displays what the student knows.

These differences help explain genre differences in the functions that the bundles were used to perform in these corpora. Based loosely based on Halliday's (1994) linguistic macrofunctions, bundles comprised these broad types:

- Research-oriented (ideational), which help writers to structure their activities and experiences of the real world (*at the beginning of*, *at the same time*, *in the present study*)
- Text-oriented (textual), concerned with the organization of the text and its elements as a message (*on the other hand*, *these results suggest that*, *in the next section*)
- Participant-oriented (interpersonal), which focus on the writer or reader of the text (*may be due to*, *it is possible that, should be noted that*)

Table 4 shows that half of all bundles related to the organization of the argument, although with considerable intergenre variation.

The relatively high proportion of text-oriented bundles in the research articles is worthy of comment. This is the most discursively crafted and rhetorically machined genre of the three, and almost two thirds of its clusters present research by engaging with a literature, providing warrants, establishing background, connecting ideas, directing readers around the text, and specifying limitations. The number of resultative markers, for example, shows a high degree of reader awareness as it points to the writer's interpretations and highlights the inferences the writer wants readers to draw:

Table 4. Distribution of Bundle Functions by Genre (%) (Hyland, 2008a: p 54)

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

(3) *The results of the* mating experiments clearly indicate the existence of two ISGs in C. subnuda. (Bio RA)
On the theoretical level, our *results suggest that the* perspective of opportunism may not axiomatically hold in all asymmetric contexts. (BS RA)

The high proportion of text-oriented bundles similarly suggest a clear audience orientation among PhD students, as well as laying claim to a certain disciplinary competence by demonstrating a care with both research and with language. Because the PhD texts were much longer, they also contained text-oriented strings which structured more discursively elaborate arguments over a greater span of text, referring to text stages and announcing discourse goals, as in item four in the following list, or pointing to other parts of the texts to make additional material salient and available to readers in recovering the writer's intentions (item 5):

(4) *In an attempt to* establish the research context for this inquiry, in section 2.5, I begin with the research history of language learner strategies and then... (AL PhD)
In this section we offer evidence on the effect of corporate investment decisions on the market value of the firm. (BS PhD)
When the system is in normal condition, the computer result *is shown in Figure* 20 and the voltage profile of the weakest bus *is shown in Figure* 21. (EE PhD)
Their styles of being a facilitator *will be discussed in* the next chapter, indicating the favorable student factors that contributed to being a facilitator. (AL PhD)

While apparently referential, these clusters have important rhetorical functions by helping to frame, scaffold, and present arguments as a coherently managed and organized arrangement. As such they reflect the writers' awareness of the discursive conventions of a sustained discussion and the processing needs of a particular disciplinary audience.

The discourse of master's students,' on the other hand, is characterized less by a text-oriented reader awareness than by use of research-oriented bundles and a relatively low use of participant-oriented forms, choices which impart a strong real-world, research-focused sense to their texts. The master's students were the only writers to refer more to their research than its presentation, drawing particularly on those clusters which described research objects or contexts (6) and, in almost 25 percent of cases, those depicting procedures (7):

(5) *The structure of the* resolver is similar to that of a motor. (EE MSc)
This is *the name of the* executable file, i.e. "winword," "excel," etc. (AL MA)

(6) Daily spiking was required *in order to maintain* the tank mercury concentration close to the designated concentration. (Bio MSc)
Parallel processing *can be used to* carry out the multistation-runs by a number of computers in order to minimize the computation time. . .. (BS MA)

Interestingly, these preferences also seem to characterize undergraduate writing among native English speakers in the British Academic Written English Corpus (Lee & Chen, 2009).

The infrequent use of participant sequences is often seen as a defining feature of expository writing by L2 students and perhaps reflects cultural preferences for a noninterventionist stance among these Hong Kong writers (Scollon & Scollon, 2001). The assertion of an explicit authorial position is, however, a common feature of published academic writing, which is clearly structured to evoke affinity and engagement. Along with their observations and interpretations, writers annotate their texts to comment on the possible accuracy of a claim, the extent they want to commitment themselves to it, or the attitude they want to convey, as here:

(7) However, this *may be due to* disruption of the complex upon antibody binding, or the antibodies we have used may block the interaction. (Bio RA)
It is obvious that the partial heat resistances are provided directly by the structure function. (EE RA)

BUNDLES AND DISCIPLINARY VARIATION

Studies show that the distribution of bundles not only characterizes particular modes, genres, and authors, but also is a strong disciplinary marker. This is clear from a disciplinary analysis of the research article, thesis, and dissertation corpora discussed earlier in this article (Hyland, 2008b). In terms of frequencies, for example, electrical engineering texts contained the greatest range of bundles with 213 different four-word strings meeting the 20 per million words threshold

Table 5. Frequency of Bundles by Discipline

Discipline	Different bundles	Total cases	% of total words in bundles
Electrical engineering	213	4562	3.5
Business studies	144	3728	2.2
Applied linguistics	141	4631	1.9
Biology	131	2909	1.7

(across 10 percent of texts) and also the highest proportion of words in the texts occurring in four-word bundles (Table 5).

Many bundles used by engineers are therefore not found in the other disciplines, and there is considerably greater reliance on prefabricated structures than in the other fields, possibly reflecting the dependence of engineering rhetoric on visual representation where formulas and graphs are linked in routinely patterned, almost formulaic ways.

There was also considerable disciplinary specificity in the four-word bundles themselves. Table 6 shows the 30 most commonly used bundles in the four fields in frequency order, with just four items occurring in all four disciplines (bolded) and a handful in three disciplines (shaded).

While *on the other hand*, *in the case of*, *as well as the,* and *at the same time* occur in each of these disciplines, different fields seem to draw on almost completely different sets of items. More than half the items in each list do not occur at all in any other discipline, and only 30 percent of the strings in each discipline are found in two other fields. Applied linguistics has 29 items in the top 50 that do not occur in any of the other lists, and electrical engineering has 28. The greatest affinity is between broadly cognate fields, as business studies and applied linguistics share 18 items, and biology and electrical engineering share 16. These contrasts perhaps 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.

A similar picture emerges with the forms and functions of these bundles. While a noun phrase with *of-* fragment is the most common structure overall, composing about a quarter of all forms in the corpus, social scientists made far greater use of bundles beginning with a prepositional phrase, typically indicating logical relations between propositional elements:

(8) 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)

This form often assists writers to discursively explore possibilities and elaborate relationships in argument. In contrast, the science and engineering

Table 6. Most Frequent 30 Four-Word Bundles in Four Disciplines (Hyland, 2008b: p 12)

Biology	Electrical engineering	Applied linguistics	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	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 other hand the
be due to the	is given by equation	in the context of	in the context of
in the case of	the effect of the	is one of the	as a result of
is shown in figure	the magnitude of the	in the process of	the performance of the
the beginning of the	**at the same time**	the results of the	is positively related to
the nature of the	in this case the	in terms of their	are significantly different from
the fact that the	it is found that	to the fact that	in terms of the
may be due to	the size of the	in the sense that	the degree to which
are summarized in table	be seen that the	the relationship between the	in the long run
has been shown to	the accuracy of the	at the beginning of	in the united states
an important role in	**as well as the**	the role of the	the nature of the
at room temperature for	the same as the	of the present study	the total number of
at the same time	is one of the	as a result of	the size of the
can be used to	a function of the	one of the most	in the number of
in the absence of	as a result the	can be seen as	it is important to
as shown in figure	the results of the	it is important to	the standard deviation of
with respect to the	in the form of	it should be noted	with respect to the
used in this study	is assumed to be	on the one hand	of the number of

texts employed about four times more passive bundles, often followed by a prepositional phrase marking a locative or logical relation, to either guide readers through the text (10) or identify the basis for an assertion (11):

(9) The experiment setup *is shown in Figure* 4.13. (EE)
All important events for pot trials *are summarized in Table* 4.11. (Bio)

(10) 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)

One major disciplinary difference in the distribution of functions is the greater concentration of research-oriented bundles in the science and engineering texts, a preference that amounted to almost half of all bundles in the science/technology corpora. Once again, this imparts a greater real-world, laboratory-focused sense to writing in the hard sciences, contributing to the description or specification of research objects or contexts:

(11) *The structure of the* coasting-point identification model (see fig 5.6) can be divided into the following areas for description. (EE)
The size of the perforations becomes progressively smaller towards *the base of the* apparatus. (Bio)

More than half of all cases, however, depicted research procedures, showing the ways that experiments and research were conducted:

(12) 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)

The social science texts, on the other hand, contained more than twice as many participant-oriented bundles as writers sought to establish their claims through more explicit evaluation and reader engagement. Here personal credibility and explicitly getting behind arguments play a far greater part in creating a convincing discourse:

(13) 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 toward separate entities. (BS)

In the sciences, participant bundles largely sought to engage readers, explicitly marking the presence of the "reader-in-the-text" (Thompson & Thetela, 1995, 103) through the use of directives (Hyland, 2002):

(14) In other words, although mixtures of zero al exists, *it is necessary to* carefully optimize the material parameters associated with the rotational viscosity. (EE)
It should be noted that the extracted MAPs are associated with the polymerized tubulin. (Bio)

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 judgment of necessity/importance.

PEDAGOGIC ISSUES

While the description of common lexical bundles can help us understand something of the features of academic writing and how disciplinary arguments are accomplished in different contexts, their study can also inform pedagogy (e.g., Meunier & Granger, 2008). Bundles are familiar to writers and readers who regularly participate in a particular discourse, their very naturalness signaling competent participation in a given community. Conversely, this means that the absence of such clusters reveal the lack of fluency of a novice or newcomer to that community. Haswell (1991), for example, suggested the following:

> 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. (p. 236)

The study of high-frequency strings and their possible variations may thus have great pedagogic value to teachers of English for academic purposes (EAP).

Research indicates, however, that the bundles used by novices and students differ markedly from those in professional academic writing (e.g., Chen & Baker, 2010; Cortes, 2004; Hyland, 2008a; Scott & Tribble, 2006). Studies have found, for example, that Chinese writers have difficulties in controlling this feature of academic writing (Lee & Chen, 2009), either overusing particular connectors, such as *first of all*, *on the other hand*, and *in a nutshell*, compared with English writers (Milton, 1998), or otherwise demonstrating a lack of fluency (Ma 2009; Wei, 2007; Xu, 2007). Schmitt et al. (2004), however, found that relatively proficient EAP learners seem to already know a considerable number of high-frequency formulaic sequences and that they enhanced this knowledge over a 3-month course. Similarly, Li and Schmitt's (2009) Chinese case study student acquired 166 new lexical phrases during her one-year MA course.

It is possible, then, for bundles to be taught in EAP classrooms, although to date very little by way of practical applications has been published. Results, moreover, have generally been mixed. While Weber (2001) was able to use concordancing methods to teach her L2 law students key lexical items which included bundles, Cortes (2006) found her short course presenting bundles

to undergraduate history students was not long enough to make a significant impact in their production. Jones and Haywood (2004), however, successfully introduced their intermediate level L2 students to frequent academic sequences in an intensive presessional EAP course. Beginning with reading texts flooded with core bundles of various lengths and using noticing activities which focused on concordance lines, the teachers then required learners to produce the sequences in cause-effect and problem-solution essays and in gapped writing tasks. Following pre- and post tests, these authors reported that through instruction and repeated exposure, "most students had shown greater awareness of formulaic sequences used as whole units, and a few students were able to us certain formulaic sequences accurately and appropriately in their essays" (Jones & Haywood, 2004, p. 290).

The recent publication of an empirically derived *Academic Formulas List* (Simpson-Vlach & Ellis, 2010) provides an impetus to further classroom practice in this area. Classified by adopting Biber, Conrad, and Cortes's (2003, 2004) pragmatic functions and identified from several academic corpora using statistical and qualitative methods, the list offers teachers a pedagogically useful inventory of sequences for speech and writing across a range of academic disciplines. While this may prove to be an important aid to instruction, work remains to be done on how best to make use of this resource.

CONCLUSIONS

Multiword expressions are an important defining feature of academic discourse and a significant component of fluent linguistic production. For these reasons there has been considerable interest in the last decade in identifying and categorizing bundles in order to characterize particular genres and harness the potential of common strings for successful language learning. Although issues of identification remain and studies suggest that corpus data on its own may be a poor indication of whether bundles are stored as chunks in the mind (Schmitt, Grandage, & Adolphs, 2004), their very frequency in academic genres testifies that they constitute an important element of scholarly rhetorical competence. The ubiquity of these features suggests that gaining control of academic discourse 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 bundles to better help learners acquire the specific rhetorical practices of their communities.

ANNOTATED BIBLIOGRAPHY

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

Chapter six has a good discussion of bundles with definitional criteria, formal and functional categories, and an analysis of textbooks and classroom teaching.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ... lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*, 371–405.

A presentation of a functionally derived classification of academic bundles.

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

A cross-genre analysis of a large corpus of academic writing distinguished by discipline.

Simpson-Vlach, R., & Ellis, N. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*, 487–512.

An empirically derived proposal for a pedagogically useful list of multiword bundles derived from spoken and written academic genres in four broad fields of inquiry; a good starting point for teaching purposes.

REFERENCES

Altenberg, B. (1993). Recurrent word combinations in spoken English. In J. D. Arcy (Ed.), *Proceedings of the Fifth Nordic Association for English Studies Conference* (pp. 17–27). Reykjavik: University of Iceland.

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word-combinations. In A. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 101–122). Oxford, UK: Oxford University Press.

Barlow, M. (2004). Collocate [Computer software]. Houston, TX: Athelstan.

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

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics, 14*, 275–311.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*, 263–286.

Biber, D., Conrad, S., & Cortes, V. (2003). Towards a taxonomy of lexical bundles in speech and writing. In A. Wilson, P. Rayson, & T. McEnery (Eds.), *Corpus linguistics by the lune: A festschrift for Geoffrey Leech* (pp. 71–92). Frankfurt, Germany: Peter Lang.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ... lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*, 371–405.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow, UK: Pearson.

Butler, C. (1998). Collocational frameworks in Spanish. *Journal of Corpus Linguistics, 3*, 1–32.

Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 student writing. *Language, learning and technology, 14*, 30–49.

Cheng, W. (2007). Concgramming: A corpus-driven approach to learning the phraseology of discipline-specific texts. *CORELL: Computer Resources for Language Learning, 1*, 22–35.

Cheng, W., Greaves, C., & Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics, 11*, 411–433.

Cheng, W., Greaves, C., Sinclair, J., & Warren, M. (2007). Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. *Applied Linguistics, 30*, 236–252.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*, 397–423.

Cortes, V. (2006). Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education, 17*, 391–406.
Cortes, V. (2008). A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora, 3*, 43–58.
Cortes, V., & Csomay, E. (2007). Positioning lexical bundles in university lectures. In M. Campoy & M. Luzon (Eds.), *Spoken corpora in applied linguistics* (pp. 55–77). New York, NY: Peter Lang.
Coxhead, A., & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing, 16*, 129–147.
De Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics, 3*, 59–80.
Ellis, N., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly, 42*, 375–396.
Greaves, C. (2009). *Concgram 1.0: A phraseological search engine*. Amsterdam, the Netherlands: John Benjamins.
Halliday, M. A. K. (1994). *Functions of language* (2nd ed.). London, UK: Edward Arnold.
Haswell, R. (1991). *Gaining ground in college writing: Tales of development and interpretation*. Dallas, TX: Southern Methodist University Press.
Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London, UK: Routledge.
Hyland, K. (2002). Directives: Argument and engagement in academic writing. *Applied Linguistics, 23*, 215–239.
Hyland, K. (2008a). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*, 41–62.
Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*, 4–21.
Hyland, K., & Milton, J. (1997). Hedging in L1 and L2 student writing. *Journal of Second Language Writing, 6*, 183–206.
Hyland, K., & Tse, P. (2005). Hooking the reader: A corpus study of *evaluative that* in abstracts. *English for Specific Purposes, 24*, 123–139.
Jones, M., & Haywood, S. (2004). Facilitating the acquisition of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences* (pp. 269–300). Amsterdam, the Netherlands: John Benjamins.
Kim, Y. (2009). Korean lexical bundles in conversation and academic texts. *Corpora, 4*: 135–165.
Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing, 18*, 85–102.
Lee, D., & Chen, S. X. (2009). Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing, 18*, 281–296.
Ma, G. H. (2009). A study of four-word lexical bundles in English major students' timed writing. *Foreign Language Teaching and Research, 41*, 54–60.
Meunier, F., & Granger, S. (2008). *Phraseology in foreign language learning and teaching*. Amsterdam, the Netherlands: John Benjamins.
Milton, J. (1998). Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment. In S. Granger (Ed.), *Learner English on computer*. London, UK: Longman.
Nesi, H., & Basturkmen, H. (2006). Lexical bundles and discourse signaling in academic lectures. *International Journal of Corpus Linguistics, 11*, 283–304.
Pang, W. (2010). Lexical bundles and the construction of an academic voice: A pedagogical perspective. *Asian EFL Journal, 47*, 1–13.
Renouf, A., & Sinclair, J. (1991). Collocational frameworks in English. In K. Aijmer & B. Altenberg (Eds.), *Advances in corpus linguistics* (128–143). Amsterdam, the Netherlands: Rodopi.

Schmitt, N., Dornyei, Z., Adolphs, S., & Durow, V. (2004). Knowledge and acquisition of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences* (55–86). Amsterdam, the Netherlands: John Benjamins.

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psychologically valid? In N. Schmitt (Ed.), *Formulaic sequences* (pp. 127–151). Amsterdam, the Netherlands: John Benjamins.

Scott, M. (2008). WordSmith Tools, Version five [Computer software]. Liverpool, UK: Lexical Analysis Software.

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

Scollon, R., & Scollon, S. (2001). *Intercultural communication* (2nd ed.). Oxford, UK: Blackwell.

Simpson, R. (2004). Stylistic features of academic speech: The role of formulaic expressions. In T. Upton & U. Connor (Eds.), *Discourse in the professions: Perspectives from corpus linguistics* (pp. 37–64). Amsterdam, the Netherlands: John Benjamins.

Simpson, R., & Mendis, D. (2003). A corpus-based study of idioms in academic speech. *TESOL Quarterly, 3*, 419–441.

Simpson-Vlach, R., & Ellis, N. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*, 487–512.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford, UK: Oxford University Press.

Swales, J. (1990). Genre analysis: English in academic and research settings. Cambridge, UK: Cambridge University Press.

Thompson, G., & Thetela, P. (2001). The sound of one hand clapping: the management of interaction in written discourse. *Text, 15*, 103–127.

Tracy-Ventura, N., Cortes, V., & Biber, D. (2007). Lexical bundles in speech and writing. In G. Parodi (Ed.), *Working with Spanish corpora* (pp. 217–230). London, UK: Continuum.

Weber, J.-J. (2001). A concordance and genre-informed approach to ESP essay writing. *ELTJ, 55*, 14–20.

Wei, N. X. (2007). Phraseological characteristics of Chinese learners' spoken English: Evidence of lexical chunks from COLSEC. *Modern Foreign Languages, 30*, 281–291.

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge, UK: Cambridge University Press.

Xu, J. J. (2007). Discourse management chunks in Chinese college learners' English speech: A spoken corpus-based study. *Foreign Language Teaching and Research, 39*, 437–443.