*Applied Linguistics 25/3: 371–405* 

# *If you look at . . .*: Lexical Bundles in University Teaching and Textbooks

DOUGLAS BIBER[1], SUSAN CONRAD[2], and VIVIANA CORTES[3]

[1]Northern Arizona University, [2]Portland State University, and [3]Iowa State University

This paper investigates the use of multi-word sequences in two important university registers: classroom teaching and textbooks. Following Biber *et al.* (1999), we take a frequency-driven approach to the identification of multi-word sequences, referred to as 'lexical bundles'. We compare the lexical bundles in classroom teaching and textbooks to those found in our previous research on conversation and academic prose. Structural patterns are described first, and then we present a functional taxonomy, including stance expressions, discourse organizers, and referential expressions. The use of lexical bundles in classroom teaching turns out to be especially surprising, both in frequency and in function. Classroom teaching uses more stance and discourse organizing bundles than conversation does, but at the same time, classroom teaching uses more referential bundles than academic prose. The analysis indicates that lexical bundles—the most frequent sequences of words in a register—are a unique linguistic construct. Lexical bundles are usually not complete grammatical structures nor are they idiomatic, but they function as basic building blocks of discourse. In the conclusion, we discuss the implications of our study for the theoretical status of lexical bundles.

## 1 INTRODUCTION

There have been many linguistic studies of academic discourse published over the past 20 years (e.g. the extensive survey of research in Grabe and Kaplan 1996). Most of these focus on the description of a specific grammatical feature or lexical class in written academic registers, especially academic research articles in science or medicine (e.g. Halliday 1988; Myers 1989; Swales 1990; Thompson and Ye 1991; Halliday and Martin 1993; Hyland 1994, 1996a, 1996b; Williams 1996; Crompton 1997; Grabe and Kaplan 1997; Swales *et al.* 1998; Chih-Hua 1999; Marco 2000; see also the survey of grammatical characteristics in Biber *et al.* 1999).

Relatively few studies describe the linguistic characteristics of spoken academic discourse. However, some of these studies take a different approach, focusing on discourse markers and other relatively fixed lexical 'chunks' instead of grammatical features (e.g. Chaudron and Richards 1986; Strodt-Lopez 1991; Nattinger and DeCarrico 1992; Flowerdew and Tauroza 1995).

These studies of discourse markers and other lexical expressions in academic lectures are part of a growing research tradition focusing on the

use of multi-word prefabricated expressions (e.g. *in a nutshell, if you see what I mean*). This research challenges the view that language is strictly compositional, arguing instead that much of our everyday language use is composed of prefabricated expressions (see the reviews in Ellis 1996; Howarth 1996; Wray and Perkins 2000; Wray 2002).

Multi-word sequences have been studied under many rubrics, including 'lexical phrases', 'formulas', 'routines', 'fixed expressions', 'prefabricated patterns' (or 'prefabs'), and 'lexical bundles'. Many previous studies have been primarily theoretical in nature, comparing the various perspectives and approaches to multi-word units, proposing new frameworks for analysis, and calling for further research. Hakuta (1974), Yorio (1980), Pawley and Syder (1983), Redeker (1991), Sinclair (1991), Lewis (1993), Weinert (1995), Howarth (1996, 1998a, 1998b), and Wray and Perkins (2000) are good examples of this type. Weinert (1995: 182) identifies two basic research issues: the best way to define and identify fixed multi-word units, and analysis of the discourse functions that these multi-word units perform.

Numerous empirical studies have addressed aspects of these two research issues (e.g. Renouf and Sinclair 1991; Kjellmer 1991; Nattinger and DeCarrico 1992; Altenberg and Eeg-Olofsson 1990; Altenberg 1998; Aijmer 1996; Francis *et al*. 1996, 1997, 1998; Butler 1997, 1998; Hudson 1998; DeCock 1998; Granger 1998; Howarth 1998a, 1998b; Moon 1998a, 1998b; Partington 1998; Hunston and Francis 1998, 1999; Gledhill 2000; Schmitt 2004). However, despite the general consensus on the importance of multi-word units, there is surprisingly little agreement on their defining characteristics, the methodologies to identify them, or even what to call them; and, as a result, there is little agreement across studies on the specific set of multi-word units worthy of description. These empirical research studies differ in terms of:

1 the research goals adopted: describing the full range of multi-word sequences vs. describing a small set of 'important' sequences;
2 the criteria used to identify multi-word units: perceptual salience, frequency criteria, or other;
3 the formal characteristics of the multi-word units studied: continuous sequences, discontinuous frames, or lexico-grammatical patterns; two-word collocations or longer sequences;
4 the text samples drawn on: ranging from a few texts to very large corpora (*c*.100 million words);
5 whether or not register comparisons are made: many studies disregard register completely; others analyse only spoken texts or written texts; a few studies explicitly compare multi-word units across different registers.

Given the complexity of these issues, we take the position that no single approach can provide the whole story. Rather, the overall importance of multi-word units in discourse can be fully understood only by undertaking empirical research studies from different perspectives. In the present paper,

we adopt a frequency perspective—'lexical bundles'—to investigate the functions of multi-word sequences in university registers.

## 1.1 Previous research on lexical bundles

The present study adopts a frequency-driven approach to multi-word units, based on analysis of the most frequent recurrent sequences of words. Salem (1987) carried out pioneering research of this type, based on analysis of a corpus of French government documents. Altenberg (1998; Altenberg and Eeg-Olofsson 1990) was probably the first researcher to study recurrent word sequences in English (based on the London-Lund Corpus), while Butler (1997) adopted a similar approach for his analysis of recurrent word sequences in a large corpus of Spanish texts. We extended this research approach in the *Longman Grammar of Spoken and Written English* (Biber *et al.* 1999, ch. 13; see also Biber and Conrad 1999), referring to these recurrent sequences of words as 'lexical bundles' (see section 2 below). That study used corpus-based research methods to compare the most common multi-word units in spoken and written registers. That research was distinctive in several respects:

- it adopted a register perspective and explicitly compared spoken and written registers (conversation and academic prose);
- it was based on empirical analysis of large corpora (*c.*5 million words for each register);
- it relied exclusively on frequency criteria for the identification of multi-word units;
- it focused on longer multi-word units than in most previous studies: 4, 5, and 6-word sequences.

In a subsequent study, we developed a preliminary taxonomy to classify the discourse functions of the lexical bundles found in conversation and academic prose (Biber *et al.* 2003; Conrad and Biber, in press). This analytical framework has also been applied to other more specialized domains. For example, Cortes (2002, to appear) applies the lexical bundle framework to the study of university student writing in history and biology, comparing the use of lexical bundles by professional and student authors. Partington and Morley (2004) investigate the use of lexical bundles in spoken political discourse (White House press briefings and political news interviews), showing how lexical bundles can be used to analyse the distinctiveness of particular registers.

## 1.2 Overview of the present study

In the present paper we turn to the use of lexical bundles in university-level courses. Specifically, we analyse the use of lexical bundles in university classroom teaching and textbooks. Classroom teaching and

textbooks are arguably the two most important registers in the academic lives of university students. However, we know surprisingly little about the language of these registers. To our knowledge, only one previous study has investigated the use of multi-word units in university lectures: the study by DeCarrico and Nattinger (1988; see also Nattinger and DeCarrico 1992) on 'lexical phrases'. The present study extends the research findings of Nattinger and DeCarrico by analysing the patterns of use in a large corpus and employing frequency criteria rather than perceptual salience as the basis for analysis.

We compare the patterns of use in classroom teaching and textbooks to those found in conversation and academic prose. Conversation can be regarded as a stereotypically 'oral' register, characterized by high interaction, expression of personal stance, and real-time production circumstances (see Biber 1995). In contrast, academic prose can be regarded as a stereotypically 'literate' register, characterized by an informational rather than personal focus, and extensive opportunity for crafting, revising, and editing the written text. As discussed in our previous work (e.g. Biber *et al*. 2002, 2004), classroom teaching and textbooks are intermediate registers with respect to these characteristics. Classroom teaching is a spoken register, constrained by real-time production circumstances, and marked to some extent by speakers' personal concerns and interactions among participants. At the same time, classroom teaching has a primary informational focus, and instructors normally pre-plan the content and structure of their class sessions to achieve their informational goals. As a result, we would expect that classroom discourse would be intermediate between conversation and academic prose in its use of lexical bundles. Textbooks are more similar to academic prose in purpose and production circumstances, but the material must be presented in a way that is accessible to students. Given recent trends to make textbooks more engaging for students, we predicted that they would also be intermediate between conversation and academic prose in their use of lexical bundles.

In section 2 below, we describe the methodology used to identify lexical bundles. Then, in section 3, we briefly summarize some of our earlier grammatical research on university registers. This research provides important background to the present study, because it raises expectations concerning the linguistic differences among spoken and written university registers; it turns out that the patterns of use for lexical bundles are strikingly different from the patterns observed for other grammatical features.

In sections 4 and 5, we turn to the analysis of lexical bundles in university registers. In section 4, we summarize our distributional findings, describing the structural types of lexical bundles found in classroom teaching and textbooks, and comparing those to the common lexical bundles found in conversation and professional academic prose. Then, in section 5 we introduce a functional taxonomy of the lexical bundles in classroom teaching and textbooks, grouping bundles into categories that serve related discourse

functions, and contrasting the use of these functional groups across spoken and written university registers.

## 2 METHODOLOGY

### 2.1 Corpus used for the study

The present study is based on an analysis of texts from university classroom teaching and textbooks in the T2K-SWAL Corpus (TOEFL 2000 Spoken and Written Academic Language Corpus; see Biber *et al*. 2002, 2004). The T2K-SWAL Corpus was designed to represent the range of spoken and written registers that university students encounter in the U.S.[1] The register categories chosen for the corpus are sampled from across a large range of spoken and written activities associated with academic life, including classroom teaching, office hours, study groups, on-campus service encounters, textbooks, course packs, and other written materials (e.g. university catalogues, brochures). The study reported here analyses only two of those registers: classroom teaching and textbooks. Table 1 shows the composition of this sub-corpus by register category.

Texts from classroom teaching and textbooks were sampled from six major academic disciplines (Business, Education, Engineering, Humanities, Natural Science, and Social Science) and three levels of education (lower division undergraduate, upper division undergraduate, and graduate). Texts were collected at four academic sites (Northern Arizona University, Iowa State University, California State University at Sacramento, and Georgia State University); the four universities are situated in different regions of the USA and have different institutional profiles. Although this sampling does not achieve complete demographic/institutional representativeness, it does avoid obvious skewing for these factors.

To provide a baseline for the analysis of university registers, we compare the patterns of use to our earlier descriptions of lexical bundles in conversation and academic prose, based on analysis of the Longman Spoken and Written English Corpus (*c*.4 million words of British English conversation; *c*.3 million words of American English conversation; *c*.5.3 million words of academic

*Table 1: Composition of the sub-corpus used in the analysis*

| Register | No. of texts | No. of words |
|---|---|---|
| Classroom teaching | 176 | 1,248,800 |
| Textbooks | 87 | 760,600 |
| Total | 263 | 2,009,400 |

**Source:** The TOEFL 2000 Spoken and Writing Academic Language—T2K-SWAL—Corpus

prose; see Biber *et al*. 1999, ch. 1). The academic prose corpus comprises both academic research articles (*c*.2.7 million words) and advanced academic books (*c*.2.6 million words; see Biber *et al*. 1999: 32–4). While some advanced academic books can also be used as textbooks, especially in graduate courses, the corpora of academic prose and textbooks were sampled independently: textbooks are mostly written specifically for students, while the articles and books included in the academic prose corpus are mostly written for other professionals.

## Identification of lexical bundles

Lexical bundles are defined simply as the most frequent recurring lexical sequences in a register (Biber *et al*. 1999, ch. 13). In the analysis of lexical bundles, as in our previous corpus-based studies of grammatical patterns, we do not regard frequency data as explanatory. In fact we would argue for the opposite: frequency data identifies patterns that must be explained. The usefulness of frequency data (and corpus analysis generally) is that it identifies patterns of use that otherwise often go unnoticed by researchers.

Frequency data have additional importance for the study of multi-word sequences because they are one reflection of the extent to which a sequence of words is stored and used as a prefabricated chunk, with higher frequency sequences more likely to be stored as unanalysed chunks than lower frequency sequences. (Of course, frequency is only one measure of the extent to which a multi-word sequence is prefabricated; sequences with idiomatic meanings are usually rare but clearly prefabricated.)

The actual frequency cut-off used to identify lexical bundles is somewhat arbitrary. For the present study, we take a conservative approach, setting a relatively high frequency cut-off of 40 times per million words to be included in the analysis. Many of the bundles described here are actually much more common, occurring more than 200 times per million words. (In contrast, we included word sequences that occurred only 10 times per million words in the Biber *et al.* (1999) study.) To further limit the scope of the investigation here, only four-word sequences are considered; for example, *do you want to* and *I don't know what* are common four-word lexical bundles in conversation. (The text excerpts in sections 4 and 5 below show that two four-word lexical bundles sometimes occur together to form a five-word or six-word sequence; see Biber *et al.* (1999: 992 ff) for discussion of these longer lexical bundles.)[2]

A sequence must be used in at least five different texts to be counted as a lexical bundle, to guard against idiosyncratic uses by individual speakers or authors. In practice, this restriction has little effect, because most bundles are distributed widely across the texts in a corpus. Even the least common lexical bundles in the present analysis (with a frequency of 40 per million words) are used in at least 20 different texts, while the more common bundles are distributed more widely.[3]

In general, most lexical bundles are not idiomatic in meaning. In fact, most

longer idioms are far too rare to be considered bundles. Stereotypical idioms such as *kick the bucket* (meaning 'die') and *a slap in the face* (meaning 'an affront') are rarely attested in natural speech or writing. When idioms and fixed formulas are used, they occur usually in fiction rather than in actual face-to-face conversation. For example, *kick the bucket* and *a slap in the face* occur around 5 times per million words in fiction, while they are rarely attested at all in actual face-to-face conversation (see Biber *et al*. 1999: 1024–6).[4]

Similarly, most lexical bundles do not represent a complete structural unit. For example, only 15 per cent of the lexical bundles in conversation can be regarded as complete phrases or clauses, while less than 5 per cent of the lexical bundles in academic prose represent complete structural units (see Biber *et al*. 1999: 993–1000). Instead, most lexical bundles bridge two structural units: they begin at a clause or phrase boundary, but the last words of the bundle are the first elements of a second structural unit. Most of the bundles in conversation bridge two clauses (e.g. *I want to know*, *well that's what I*), while bundles in academic prose usually bridge two phrases (e.g. *in the case of*, *the base of the*).

Our research approach to the study of lexical bundles is deliberately exploratory. We start out by simply asking whether there are chunks of language—sequences of words—that are used repeatedly by speakers and writers. The answer to this question is 'yes': there are many lexical bundles used with high frequency in texts, and it further turns out that different registers tend to rely on different sets of lexical bundles. These distributional facts raise a second set of research questions: what are these word chunks, what are their structural and functional characteristics, and how can we explain the repeated use of these extended word chunks? For the most part, linguists have not noticed these high frequency multi-word sequences, probably because most previous research has focused on grammatical phrases and clauses, disregarding the possibility of lexical units that cut across grammatical structures. However, we show below that lexical bundles have identifiable discourse functions, suggesting that they are an important part of the communicative repertoire of speakers and writers, even though they do not correspond to the well-formed structures traditionally recognized by linguistic research. We return to discussion of the theoretical status of lexical bundles in section 6.

## 3 BACKGROUND: GRAMMATICAL DIFFERENCES AMONG SPOKEN AND WRITTEN UNIVERSITY REGISTERS

In terms of their grammatical characteristics, spoken university registers are consistently different from written registers (see Biber *et al*. 2002, 2004). Two general patterns are found: (1) spoken registers as a group are different from written registers, but (2) there are small differences among spoken registers,

forming a cline from the more conversational registers to the more informational registers like classroom teaching.

Figure 1 illustrates these patterns, plotting the use of nouns, verbs, and personal pronouns in the four registers considered in the present study: conversation, classroom teaching, textbooks, and academic prose.[5] The figure shows a major difference between spoken and written registers in the use of these features: nouns are much more common in the written registers than spoken registers, while verbs and pronouns show the opposite distribution, being much more common in the spoken registers. Nouns are slightly more common in classroom teaching than in conversation, reflecting to some extent the primary informational purposes of teaching in contrast to the interpersonal purposes of conversation.

The most surprising finding to emerge from these previous grammatical studies is the extent to which classroom teaching is similar to conversation, rather than combining the linguistic characteristics of speech and informational writing. (The Multi-Dimensional analysis presented in Biber *et al.* (2002) shows these same patterns with respect to a much wider range of linguistic features.) This pattern of use contradicts our prior expectations. Given that classroom teaching is clearly pre-planned and has a primary informational focus, we expected that it would use grammatical features in a similar way to written registers like textbooks. Instead, we found that classroom teaching was most strongly influenced by its 'oral' situational characteristics—the real-time production circumstances, the interactions among participants, and a focus on the speakers' personal concerns.

Given the importance of the speech/writing distinction for grammatical

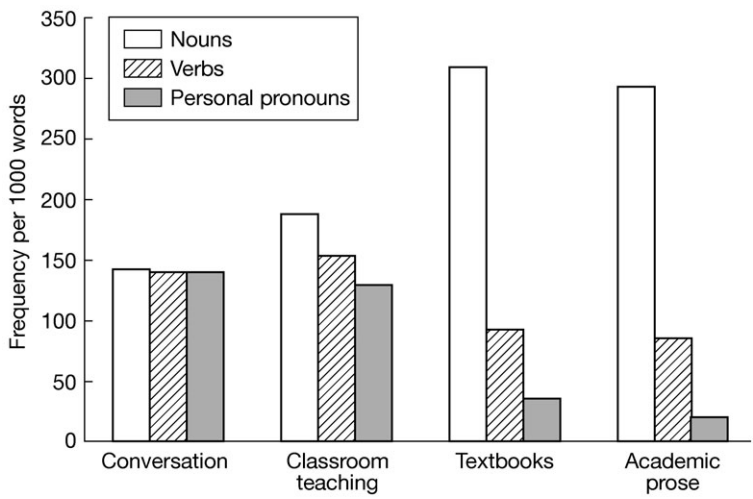


*Figure 1: Distribution of nouns, verbs, and personal pronouns across registers*

features, we expected that this difference would be equally important for the use of lexical features. Thus, we predicted that classroom teaching and conversation would be sharply distinguished from textbooks and academic prose in the use of lexical bundles. However, as the following sections show, these registers have surprisingly different patterns of use for lexical bundles. In particular, classroom teaching incorporates both 'oral' lexical bundles and 'literate' lexical bundles, in contrast to its reliance on stereotypically spoken patterns of use for grammatical features.

## 4 LEXICAL BUNDLES IN UNIVERSITY CLASSROOM TEACHING AND TEXTBOOKS

As can be seen in Figure 2, there are 43 lexical bundles in conversation; 84 in classroom teaching; 27 in textbooks; and 19 in academic prose. Overall, then, Figure 2 shows that the two spoken registers use a much greater range of different lexical bundles than the written registers. However, the most surprising finding here is for classroom teaching: rather than being intermediate between conversation and the written registers, classroom teaching far exceeds conversation in the number of different lexical bundles. Figure 3 shows that lexical bundles occur most frequently in classroom teaching as well, although they are almost as frequent in conversation. Taken together, Figures 2 and 3 show that classroom teaching uses a large set of different lexical bundles, while conversation relies on the extremely frequent use of a smaller set of bundles. For example, conversation uses 44 high-frequency bundles (occurring more than 60 times per million words). In contrast, classroom teaching has only 28 bundles that occur more than 60 times per million words.

We can begin to explain these patterns by considering the structural

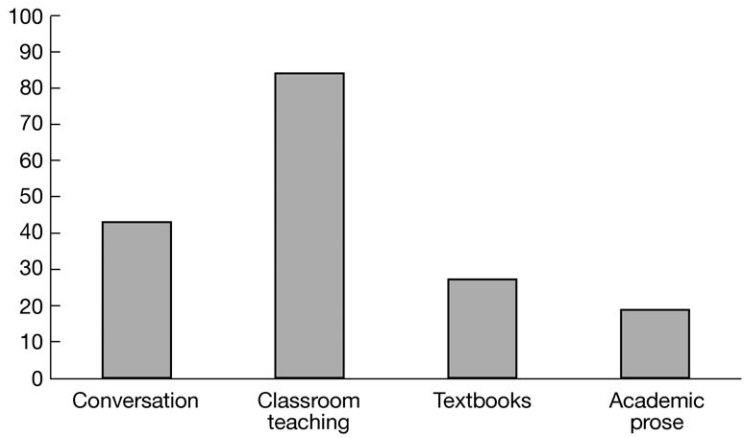


*Figure 2: Number of different leixal bundles across registers*

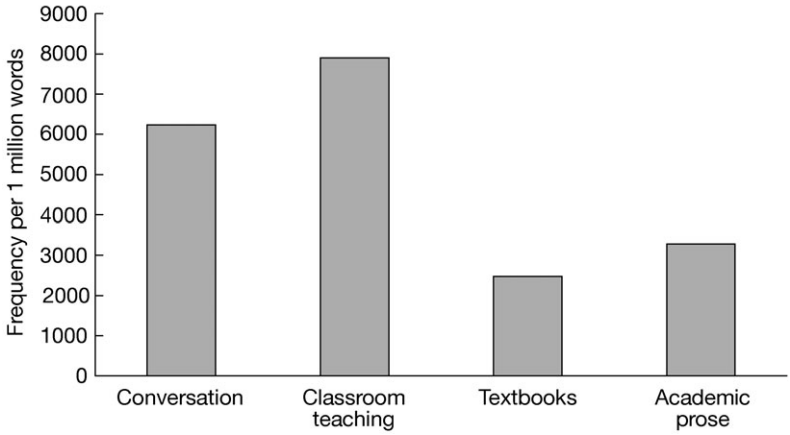


*Figure 3: Overall frequency of lexical bundles across registers*

characteristics of lexical bundles found in each register. We noted in section 2 that lexical bundles have strong grammatical correlates, even though they are not usually complete structural units. In the present study, we identified three major structural types of lexical bundle, shown in Table 2.

Type 1 bundles incorporate verb phrase fragments. For example, Types 1a and 1b begin with a subject pronoun followed by a verb phrase (e.g. *it's going to be*). Type 1c begins with a discourse marker followed by a verb phrase (e.g. *you know this is*). Types 1d and 1e simply begin with a verb phrase (e.g. *is going to be*), while Types 1f and 1g are question fragments.

Type 2 bundles incorporate dependent clause fragments in addition to simple verb phrase fragments. Type 2a bundles (e.g. *I want you to*) begin with a main clause followed by a complementizer (e.g. *to, if*) or a *WH*-word introducing a dependent clause. Other Type 2 bundles are dependent clause fragments beginning with a complementizer or subordinator (e.g. *to be able to*, *if we look at*, *what I want to*).

In contrast to Types 1 and 2, which have clausal components, Type 3 bundles are phrasal. Types 3a–3c consist of noun phrase components, usually ending with the start of a postmodifier (e.g. *the end of the*, *those of you who*). Type 3d consists of prepositional phrase components with embedded modifiers (e.g. *of the things that*), while Type 3e incorporates comparative expressions.

Figure 4 shows the distribution of these structural types across registers. Our earlier research on conversation and academic prose showed that the grammatical correlates of lexical bundles in conversation are strikingly different from those in academic prose. In conversation, almost 90 per cent of all common lexical bundles incorporate verb phrases. In fact, approximately 50 per cent of these lexical bundles begin with a personal pronoun + verb phrase (such as *I was going to*, *I thought that was*). An additional 19 per cent of

*Table 2: Structural types of lexical bundles*

| | |
|---|---|
| 1. | Lexical bundles that incorporate *verb phrase* fragments |
| 1a. | (connector +) 1st/2nd person pronoun + VP fragment<br>Example bundles: *you don't have to, I'm not going to, well I don't know* |
| 1b. | (connector +) 3rd person pronoun + VP fragment<br>Example bundles: *it's going to be, that's one of the, and this is a* |
| 1c. | Discourse marker + VP fragment<br>Example bundles: *I mean you know, you know it was, I mean I don't* |
| 1d. | Verb phrase (with non-passive verb):<br>Example bundles: *is going to be, is one of the, have a lot of, take a look at* |
| 1e. | Verb phrase with passive verb:<br>Example bundles: *is based on the, can be used to, shown in figure N* |
| 1f. | *yes–no* question fragments:<br>Example bundles: *are you going to, do you want to, does that make sense* |
| 1g. | WH-question fragments:<br>Example bundles: *what do you think, how many of you, what does that mean* |
| 2. | Lexical bundles that incorporate *dependent clause* fragments |
| 2a. | 1st/2nd person pronoun + dependent clause fragment<br>Example bundles: *I want you to, I don't know if, I don't know why, you might want to* |
| 2b. | WH-clause fragments:<br>Example bundles: *what I want to, what's going to happen, when we get to* |
| 2c. | *If*-clause fragments:<br>Example bundles: *if you want to, if you have a, if we look at* |
| 2d. | (verb/adjective+) *to*-clause fragment<br>Example bundles: *to be able to, to come up with, want to do is* |
| 2e. | *That*-clause fragments:<br>Example bundles: *that there is a, that I want to, that this is a* |
| 3. | Lexical bundles that incorporate *noun phrase and prepositional phrase* fragments |
| 3a. | (connector +) Noun phrase with *of*-phrase fragment:<br>Example bundles: *one of the things, the end of the, a little bit of* |
| 3b. | Noun phrase with other post-modifier fragment:<br>Example bundles: *a little bit about, those of you who, the way in which* |
| 3c. | Other noun phrase expressions:<br>Example bundles: *a little bit more, or something like that, and stuff like that* |
| 3d. | Prepositional phrase expressions:<br>Example bundles: *of the things that, at the end of, at the same time* |
| 3e. | Comparative expressions:<br>Example bundles: *as far as the, greater than or equal, as well as the* |

the bundles consist of an extended verb phrase fragment (e.g. *have a look at*), while another 17 per cent of the bundles are question fragments (e.g. *can I have a*). In contrast, the lexical bundles in academic prose are phrasal rather than clausal. Almost 70 per cent of the common bundles in academic prose consist of noun phrase expressions (e.g. *the nature of the*) or a sequence that bridges across two prepositional phrases (e.g. *as a result of*).

Classroom teaching uses about twice as many different lexical bundles as conversation, and about four times as many as textbooks. The distribution across structural patterns shown in Figure 4 helps explain this extremely frequent pattern of use. In marked contrast to the general patterns of use for grammatical features (described in section 3 above), classroom teaching relies on the lexical bundles associated with both spoken and written registers. Similar to conversation, classroom teaching makes dense use of lexical bundles that represent declarative and interrogative clause fragments. At the same time, classroom teaching is similar to academic prose and textbooks in making dense use of noun phrase and prepositional phrase lexical bundles. Thus, the extremely high density of lexical bundles in classroom teaching exists because this register relies heavily on both 'oral' and 'literate' bundles.

In addition, classroom teaching has a large inventory of lexical bundles associated with dependent clause fragments, especially conditional adverbial clauses and complement clauses (26 different bundles in classroom teaching, versus 16 in conversation, and only two in academic prose, and two in textbooks). This pattern is surprising, given previous claims that dependent clauses are more typical of written prose than speech (e.g. O'Donnell 1974; Kroll 1977; Chafe 1982; Akinnaso 1982; Gumperz *et al*. 1984). However, it turns out that these lexical bundles are more common in both classroom teaching and in conversation than in the written registers (similar to the

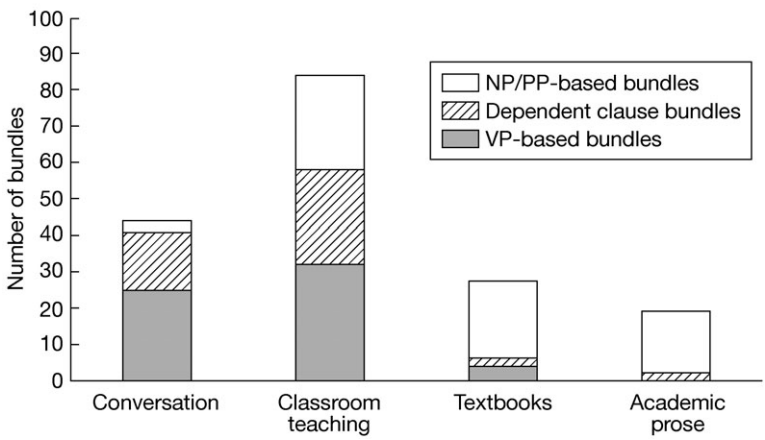


*Figure 4: Distribution of lexical bundles across structural types*

general grammatical patterns of use for verb+complement clause constructions; see Biber *et al*. 1999, ch. 9). We return to a discussion of these bundles in section 5.5 below.

Textbooks and academic prose are at the opposite extreme from classroom teaching on Figures 2 and 3.[6] It is surprising that textbook authors do not incorporate more lexical bundles in their writing, given the heavy reliance on bundles in classroom teaching. Reasons for this absence might be that textbook authors tend to use fuller expressions, preferring full clauses rather than phrasal lexical bundles, as well as the fact that textbook authors are free of the real-time production constraints of face-to-face teaching and therefore make more diverse language choices. At the present stage of our analysis, we are able only to note this distributional difference between textbooks and classroom teaching. A much fuller analysis of textbook discourse structures is required to interpret the relative absence of lexical bundles in that register.

Given that the structural correlates of lexical bundles in these four registers are so dramatically different, it will come as no surprise that their typical discourse functions differ as well. We turn to a discussion of those functions in the following section.

## 5 A PRELIMINARY FUNCTIONAL TAXONOMY OF THE LEXICAL BUNDLES IN UNIVERSITY CLASSROOM TEACHING AND TEXTBOOKS

In the present section, we turn to the discourse functions of lexical bundles, revising and extending our earlier functional taxonomy of lexical bundle types developed for conversation and academic prose (Biber *et al*. 2003). For this analysis, we examined concordance listings to analyse the functions of common bundles in their extended discourse contexts, focusing especially on the lexical bundles used in classroom teaching and textbooks. We have two immediate goals in doing this analysis: (1) to develop a taxonomy that captures the major discourse functions served by lexical bundles, and (2) to describe the extent to which each register uses lexical bundles for each of these functions.

Our approach to developing the functional taxonomy is primarily inductive. That is, we group together bundles that serve similar functions, based on the typical meanings and uses of each bundle. We used concordance listings to examine the use of each bundle in its discourse contexts. Once the bundles were assigned to groups, we attempted to determine the discourse functions associated with each of the groups. Of course this last step was influenced by previous theoretical studies on the discourse functions of linguistic features (e.g. Hymes 1974: 22ff; Halliday 1978; Brown and Fraser 1979; Biber 1988: 33ff, 1995).

In some cases, a single bundle has multiple functions even in a single occurrence. For example, bundles like *take a look at* and *let's have a look* function as both directives and topic introducers. In other cases, a single

bundle serves different functions depending on the context. For example, bundles like *the beginning of the* and *at the end of* can function as a time reference, place reference, or text deictic reference. In general, however, most bundles have a primary function. We have examined potentially multi-functional bundles in concordance listings and classified them according to their most common use. (Several of these cases are discussed in 5.1–5.3 below.)

We distinguish among three primary functions served by lexical bundles in these registers: (1) stance expressions, (2) discourse organizers, and (3) referential expressions. Stance bundles express attitudes or assessments of certainty that frame some other proposition. Discourse organizers reflect relationships between prior and coming discourse. Referential bundles make direct reference to physical or abstract entities, or to the textual context itself, either to identify the entity or to single out some particular attribute of the entity as especially important. Each of these categories has several sub-categories associated with more specific functions and meanings. Table 3 lists the bundles grouped into each of these functional categories, showing the distribution of each bundle across registers. The following subsections describe each of these categories in detail.

*Table 3: Functional classification of common lexical bundles across registers*

| | Classroom teaching | Textbooks | Conversation | Academic prose |
|---|---|---|---|---|
| I. STANCE EXPRESSIONS | | | | |
| **A. Epistemic stance** | | | | |
| *Personal*: | | | | |
| I don't know if | ** | | ** | |
| I don't know what | ** | | *** | |
| I don't know how | ** | | ** | |
| I don't know I | ** | | | |
| and I don't know | ** | | . | |
| I think it was | ** | | ** | |
| and I think that | ** | | | |
| you know what I | ** | | ** | |
| I don't think so | * | | ** | |
| I thought it was | * | | ** | |
| well I don't know | | | ** | |
| I don't know whether | | | ** | |
| I don't know why | | | ** | |
| oh I don't know | | | ** | |
| *Impersonal*: | | | | |
| are more likely to | | ** | | . |
| the fact that the | . | ** | | ** |

| | Classroom teaching | Textbooks | Conversation | Academic prose |
|---|---|---|---|---|
| **B. Attitudinal/modality stance** | | | | |
| B1) desire | | | | |
| *Personal*: | | | | |
| if you want to | *** | | ** | |
| I don't want to | ** | | *** | |
| do you want to | * | | *** | |
| you want to go | | | ** | |
| do you want a | | | ** | |
| what do you want | | | ** | |
| B2) obligation/directive | | | | |
| *Personal*: | | | | |
| I want you to | *** | | | |
| you don't have to | ** | | ** | |
| you don't want to | ** | | ** | |
| you have to be | ** | | . | |
| you have to do | ** | | * | |
| you look at the | ** | | | |
| you might want to | ** | | | |
| you need to know | ** | | | |
| and you have to | ** | | . | |
| going to have to | ** | | ** | |
| you want me to | | | ** | |
| do you want me | | | ** | |
| *Impersonal*: | | | | |
| it is important to | | ** | | * |
| it is necessary to | | | | ** |
| B3) Intention/prediction | | | | |
| *Personal*: | | | | |
| I'm not going to | ** | | ** | |
| we're going to do | ** | | | |
| we're going to have | ** | | * | |
| and we're going to | ** | | | |
| I was going to | * | | *** | |
| what we're going to | ** | | | |
| are we going to | * | | ** | |
| are you going to | * | | *** | |
| *Impersonal*: | | | | |
| it's going to be | ** | | ** | |
| is going to be | *** | | | |
| are going to be | ** | | | |
| going to be a | ** | | ** | |
| going to be the | ** | | | |

*Table 3: cont.*

| | Classroom teaching | Textbooks | Conversation | Academic prose |
|---|---|---|---|---|
| not going to be | ** | | * | |
| going to have a | * | | ** | |
| B4) ability | | | | |
| *Personal*: | | | | |
| to be able to | *** | . | . | * |
| to come up with | ** | | | |
| *Impersonal*: | | | | |
| can be used to | | ** | | * |
| it is possible to | | * | | ** |
| II. DISCOURSE ORGANIZERS | | | | |
| **A. Topic introduction/focus** | | | | |
| what do you think | ** | | ** | |
| if you look at | ** | | | |
| take a look at | ** | | | |
| if you have a | ** | | | |
| if we look at | ** | | | |
| going to talk about | ** | | | |
| to look at the | ** | | | |
| to go ahead and | ** | | | |
| I want to do | ** | | * | |
| what I want to | ** | | . | |
| want to do is | ** | | | |
| want to talk about | ** | | | |
| you know if you | ** | | | |
| a little bit about | ** | | | |
| I would like to | * | | ** | |
| in this chapter we | | ** | | |
| I/I'll tell you what | | | ** | |
| have a look at | | | ** | |
| let's have a look | | | ** | |
| do you know what | | | ** | |
| **B. Topic elaboration/clarification** | | | | |
| has to do with | ** | . | | |
| to do with the | ** | . | . | . |
| I mean you know | ** | | . | |
| you know I mean | ** | | . | |
| nothing to do with | * | . | ** | |
| on the other hand | ** | ** | | *** |
| as well as the | . | ** | | * |
| know what I mean | | | ** | |

| | Classroom teaching | Textbooks | Conversation | Academic prose |
|---|---|---|---|---|
| was going to say | | | ** | |
| what do you mean | | | ** | |
| III. REFERENTIAL EXPRESSIONS | | | | |
| **A. Identification/focus** | | | | |
| that's one of the | ** | | | |
| and this is a | ** | | | |
| and this is the | ** | | | |
| is one of the | ** | ** | | * |
| was one of the | ** | . | | |
| one of the things | *** | | | |
| and one of the | ** | | | |
| one of the most | * | | ** | ** |
| those of you who | ** | | | |
| of the things that | *** | | | |
| **B. Imprecision** | | | | |
| or something like that | ** | | ** | |
| and stuff like that | ** | | | |
| and things like that | ** | | * | |
| **C. Specification of attributes** | | | | |
| C1) Quantity specification | | | | |
| there's a lot of | ** | | * | |
| have a lot of | ** | | | |
| and a lot of | ** | | | |
| a lot of people | ** | | | |
| a lot of the | ** | | | |
| how many of you | ** | | | |
| in a lot of | ** | | | |
| the rest of the | ** | ** | * | * |
| a little bit of | ** | | | |
| a little bit more | ** | | | |
| a lot of times | ** | | | |
| than or equal to | ** | . | | |
| greater than or equal | ** | | | |
| per cent of the | | | | ** |
| C2) Tangible framing attributes | | | | |
| the size of the | . | ** | | * |
| in the form of | . | ** | | ** |
| C3) Intangible framing attributes | | | | |
| the nature of the | . | ** | | ** |
| in the case of | ** | ** | | *** |
| in terms of the | ** | ** | | * |

*Table 3: cont.*

| | Classroom teaching | Textbooks | Conversation | Academic prose |
|---|---|---|---|---|
| as a result of | * | ** | | ** |
| on the basis of | . | ** | | ** |
| in the absence of | | * | | ** |
| the way in which | * | * | | ** |
| the extent to which | | * | | ** |
| in the presence of | | | | ** |
| **D. Time/place/text reference** | | | | |
| D1) Place reference | | | | |
| the United States and | * | ** | | |
| in the United States | ** | *** | | * |
| of the United States | * | ** | | |
| D2) Time reference | | | | |
| at the same time | ** | ** | * | ** |
| at the time of | . | * | | ** |
| D3) Text deixis | | | | |
| shown in figure N | | ** | | * |
| as shown in figure | | ** | | * |
| D4) Multi-functional reference | | | | |
| the end of the | ** | ** | ** | ** |
| the beginning of the | * | ** | | * |
| the top of the | * | ** | . | . |
| at the end of | ** | *** | ** | ** |
| in the middle of | ** | * | * | |
| IV. SPECIAL CONVERSATIONAL FUNCTIONS | | | | |
| **A. Politeness** | | | | |
| thank you very much | | | ** | |
| **B. Simple inquiry** | | | | |
| what are you doing | | | ** | |
| **C. Reporting** | | | | |
| I said to him/her | | | ** | |

Key to symbols:
. = 10–19 per million words
* 20–39 per million words
** = 40–99 per million words
*** = over 100 per million words

## 5.1 Stance bundles

Stance bundles provide a frame for the interpretation of the following proposition, conveying two major kinds of meaning: epistemic and attitude/modality. Epistemic stance bundles comment on the knowledge status of the information in the following proposition: certain, uncertain, or probable/possible (e.g. *I don't know if*, *I don't think so*). Attitudinal/Modality stance bundles express speaker attitudes towards the actions or events described in the following proposition (e.g. *I want you to*, *I'm not going to*). Stance bundles can be personal or impersonal. Personal stance bundles are overtly attributed to the speaker/writer, as in the examples given above. Impersonal stance bundles express similar meanings without being attributed directly to the speaker/writer (e.g. *it is possible to*, *can be used to*).

### 5.1.1 Epistemic stance bundles

*Personal epistemic bundles*: Most epistemic stance bundles are personal. Although epistemic stance bundles can express certainty or uncertainty, most of these bundles express only uncertainty, as in:

> That's, kind of hard to tell, but again, the important thing is be resourceful when you do these. *I don't know what*, *I don't know what* the voltage is here, so, but, the real point is it's irrelevant. (classroom teaching)

Expressions with *I (don't) think* express possibility but a lack of certainty. Thus compare the two bundles in the following:

> *I don't know if* it will mean revolution in the same sense of the word, *I don't think so* because I think there are other political factors involved. (classroom teaching)

Several lexical bundles in classroom teaching combine epistemic stance with other functions. For example, bundles with *I think/thought it* serve a dual function of referential identification (see III.A in Table 3, discussed in 5.3 below) combined with an uncertain epistemic stance; for example:

> The Wall Street Journal last week or *I think it was* the Wall Street Journal had something about NASA and this same problem. (classroom teaching)

Imprecision bundles like *and stuff like that*, discussed in 5.3 below, also serve an epistemic function combined with referential identification. The bundle *what do you think* functions to introduce a new topic but also expresses an epistemic stance (see 5.2 below).

*Impersonal epistemic bundles*: In contrast to the personal epistemic bundles, the impersonal epistemic stance bundles express degrees of certainty rather than uncertainty.

> Boys *are more likely to* be hyperactive, disruptive, and aggressive in class. (textbook)

> Yet there was irony *in the fact that the* Russian Revolution, one of the most important Western revolutions, proclaimed itself to be Marxist in aims and character but happened in violation of Marxist historical logic. (textboook)

### 5.1.2 Attitudinal/modality stance bundles

Attitudinal/modality stance bundles are also usually personal, expressing speaker attitudes towards the actions or events described in the following proposition. Four major subcategories are distinguished here: desire, obligation/directive, intention/prediction, and ability.

*Desire bundles*: Desire bundles include only personal expressions of stance, which frame self-motivated wishes and desires or inquire about another participant's desires:

> So I may not want to see her face to face because *I don't want to* deliver bad news to her. (classroom teaching)

Several lexical bundles that express personal desire in classroom teaching are also used to initiate new topics, including *what I want to do* and *I would like to*; these are discussed in section 5.2 below.

*Obligation/directive bundles*: The second subcategory of attitudinal/modality stance bundles expresses obligations or directives. Most of these bundles are personal stance expressions, but they differ from other personal bundles in that they have a second person pronoun (*you*) rather than first person pronoun as subject. However, they are still clearly understood as personal expressions of stance, directing the listener to carry out actions that the speaker wants to have completed. For example:

> Now *you need to know* how to read these. (classroom teaching)

> All *you have to do* is work on it. (classroom teaching)

In some cases, these bundles include verbs of desire with a first person pronoun, directly conveying the speaker's desire that the addressee carry out some action, and thus functioning as directives:

> *I want you to* take out a piece of paper and jot some notes down in these four areas. (classroom teaching)

In other cases, the directive force of some of these bundles can be very indirect, as in:

> *You might want to* look at a couple of examples just to remind yourself of how these look. (classroom teaching)

Several bundles can fit either the desire category or obligation/directive, depending on the context of use. For example, the following would be a desire

bundle if it truly is a question about the participant's preference, but it is more likely to function as a directive when said by a teacher to a student:

> So, *do you want to* be next Erin? (classroom teaching)

Some directive bundles are used for topic introduction (e.g. *take a look at*); these are discussed in 5.2 below.

A few obligation/directive stance bundles are impersonal, with no personal pronoun at all, even though they still clearly direct the reader to carry out some action:

> *It is important to* note that Derrida does not assert the possibility of thinking outside such terms. (textbook)

*Intention/prediction bundles*: The third subcategory of attitudinal/modality stance bundles is intention/prediction. Many of these bundles are overtly personal, expressing the speaker's own intention to perform some future action. In most cases, these are expressions of joint action, used to announce the proposed plan of a class session, as in:

> But, right now *what we're going to* take a look at are ones that are produced that are positive and beneficial. (classroom teaching)

Other bundles in this category are impersonal, expressing predictions of future events that do not entail the volition of the speaker. These bundles are usually used when explaining a logical or mathematical process that involves several steps, as in:

> And so if you require a, twenty percent return on investment, this net present value *is going to be* zero. (classroom teaching)

*Ability bundles*: Only four attitudinal/modality stance bundles express ability. In their form, all four bundles are impersonal, although these bundles sometimes occur together with directive bundles to identify skills and tasks that students should accomplish:

> I want you *to be able to* name and define those four curriculum category [sic]. (classroom teaching)

> So encoding's always harder than decoding. Cos you have *to come up with* the word, you have to spell it, you have to use it correctly. (classroom teaching)

## 5.2 Discourse organizing bundles

Discourse organizing bundles serve two major functions: topic introduction/focus and topic elaboration/clarification.

### 5.2.1 Topic introduction/focus bundles

Topic introduction bundles in classroom teaching provide overt signals to the student that a new topic is being introduced. Many of these are expressions of

intention or desire (see 5.1 above), but they have the more specialized function of announcing the instructor's intention to begin a new topic:[7]

> But, before I do that, *I want to talk about* Plato. (classroom teaching)
>
> *What I want to do is* quickly run through the exercise that we're going to do. OK just so you see what it does. (classroom teaching)

As the preceding example shows, sometimes two four-word bundles occur together, in effect creating a longer five-word or six-word bundle.

The following example illustrates the use of these longer bundles for procedural instructions, identifying the major steps in the procedure:

> OK? Next thing *I want to do is—what I want to do is* I want to change the back color [. . .] OK? First thing *I want to do is* let's set up some colors of the text boxes to start with [. . .] OK? First thing *I want to do is* let's make the first text box. (classroom teaching)

Examples like this provide strong evidence to support the claim that lexical bundles are stored as unanalysed units in the mental lexicon. These topic introducing bundles often result in 'syntactic blends' (see Biber *et al*. 1999: 1064–6): the bundle signals that a new topic (or step in the procedure) is coming up, followed by the statement of the topic, but the two parts are not well-formed syntactically. For example, the bundle used repeatedly in the excerpt above—*I want to do is*—requires an obligatory subject predicative to make it syntactically complete. However, what we find instead is that the speaker starts over with a new syntactic construction (a *let's* imperative in the last two occurrences). The bundle here serves the discourse function of identifying a new topic, and it therefore functions as if it were syntactically complete; the following constituent begins a new main clause, rather than completing the subject predicative structure required by the bundle.

Topic introducing bundles can occur with both first and second person pronouns. The first person plural pronoun *we* as subject seems to invite student participation, although the 'we' often refers to the instructor rather than a collective enterprise:

> Today we are *going to talk about* testing hypotheses. (classroom teaching)
>
> Now, we *want to talk about* getting our sample mean . . . (classroom teaching)

Topic introducing bundles with second person pronouns invite student participation, although the instructor is usually intending collective consideration of the topic:

> *If you look at* development and the jobs that are created, it says nothing first of all of the type of jobs that are created. (classroom teaching)

The bundle *if you look at* often has a deictic reference, identifying the props required for a topic. The bundle directs students' attention to the prop, indirectly introducing a new topic by reference to it:

> *If you look at* the answers that are given, there's only two answers that have one big M . . . (classroom teaching)

Finally, topic introducing bundles with WH-question structures provide the most overt attempts to directly engage students in a new topic:

> *What do you think* the text is trying to tell us when they call our attention that often conflict doesn't appear suddenly? (classroom teaching)

### 5.2.2 Topic elaboration/clarification bundles

The second major subcategory of discourse organizing bundles relates to topic elaboration or clarification. For example,

> Well why is the Navajo Depot, Camp Navajo important today? [. . .] It *has to do with the* START talks with the Russians, the START Treaty signed in 1991. (classroom teaching)

The discourse markers *you know* and *I mean* are used in sequence as a lexical bundle, usually when the speaker believes that additional explanation or clarification is required:

> When you come to class next time—and I'm gonna look at grammar *you know I mean* I expect you to have things spelled relatively correctly . . . (classroom teaching)

The bundles *as well as the* and *on the other hand* are used for explicit comparison and contrast. These two discourse organizing bundles are considerably more common in textbooks than in classroom teaching:

> Section 3.5 [. . .] illustrates how the techniques are employed together *as well as the* range of resulting execution characteristics that are presented to an architecture, . . . (textbook)

> We know that if the project is in the same line business as the firm's other projects [. . .] then high stand alone risk translates into high corporate risk [. . .] *On the other hand*, if the project is not in the same line business, then it is possible that the correlation may be low . . . (textbook)

## 5.3 Referential bundles

Referential bundles generally identify an entity or single out some particular attribute of an entity as especially important. We describe four major sub-categories included under referential bundles: identification/focus,

imprecision indicators, specification of attributes, and time/place/text reference.

### 5.3.1 Identification/focus bundles

Identification/focus bundles are common in classroom teaching, focusing on the noun phrase following the bundle as especially important. For example, the bundle *those of you who* identifies the subgroup of students who are in focus:

> For *those of you who* came late I have the, uh, the quiz. (classroom teaching)

In many cases, identification/focus bundles also have a discourse organizing function (see 5.2). These bundles are often used after a lengthy explanation to emphasize or summarize the main point:

> Schizophrenia typically uh will mean that uh separation from reality uh it can mean uh uh you know extreme periods of euphoria and extreme periods of depression it can mean a lot of things—and *that's one of the* problems of schizophrenia. (classroom teaching)

> OK. Uh we create a tri-block for an object of type thread, and there is a built-in thread object that has a method called sleep, and that method called sleep takes a parameter which is the number of milliseconds [. . .] OK? *And this is a* real simple way, the simplistic way to do animation. (classroom teaching)

In other cases, identification/focus bundles can be used to introduce a discussion by stating the main point first, and then giving the details:

> *One of the things* they stress in parenting is to be consistent and particularly with parents um some parents are inconsistent between siblings. Uh fathers are notorious for letting their little darling girls get away with what they swat the boys about . . . (classroom teaching)

### 5.3.2 Imprecision bundles

A second major subcategory of referential bundles indicates imprecise reference. These have two specific functions, either to indicate that a specified reference is not necessarily exact, or to indicate that there are additional references of the same type that could be provided:

> I think really we now have what about, six weeks left in class *or something like that*. (classroom teaching)

> There are obviously companies that do uh evaluations *and things like that* (classroom teaching)

### 5.3.3 Bundles specifying attributes

The third subcategory of referential bundles identify specific attributes of the following head noun. Some of these bundles specify quantities or amounts:

> You'd *have a lot of* power. (classroom teaching)
>
> Does it create a lot of wealth? No. It creates *a little bit of* wealth. (classroom teaching)

The bundle *a little bit about* usually has a more specialized function to introduce a topic (see 5.2 above), apparently to minimize the expectations required from students:

> So I want to talk *a little bit about* process control from that point of view. (classroom teaching)

Other bundles in this category describe the size and form of the following head noun:

> These figures give an idea of *the size of the* ethnological community in Russia. (textbook)
>
> They are *in the form of* half-wheels, with concentric bands of representations alternating with bands of scrollwork. (textbook)

In contrast, some specifying bundles identify abstract characteristics:

> Rather than reading textbooks and solving textbook problems, students must define and constantly refine *the nature of the* problem . . . (textbook)

These abstract specifying bundles are often used to establish logical relationships in a text:

> Fleshy fruits are classified *on the basis of* the differentiation of the fruit wall (pericarp). (textbook)
>
> They are defined *in terms of the* emotion they elicit. (textbook)

### 5.3.4 Time/place/text-deixis bundles

Finally, several referential bundles refer to particular places, times, or locations in the text itself. Three place bundles in our corpus refer to the United States, reflecting one focus of textbooks and classroom teaching in our corpus:

> Children *in the United States* are not formally employed in farm work, . . . (textbook)

Text deixis bundles are common only in the written registers, where they make direct reference to figures contained in the text itself:

> *As shown in Figure* 4.4, the higher the real estate agents scored in terms of the proactive personality dimension, the more houses they sold . . . (textbook)

Many of these bundles are multi-functional, referring to a place, time, and/or text deixis, depending on the particular context:

> So you have to record that, since the asset was sold at *the end of the* year (classroom teaching)
>
> She's in that.. uh.. office down there.. at *the end of the* hall (classroom teaching)
>
> uh I'm going to start actually with *the end of the* chapter (classroom teaching)

## 5.4 Register variation in the functional exploitation of lexical bundles

In the preceding sections, we have outlined a taxonomy of the major discourse functions served by lexical bundles. The taxonomy was developed to include functions that can potentially be realized in any register. However, as Table 4 shows, the four registers show dramatic differences in their reliance on particular functional types. The examples presented in sections 5.1–5.3 illustrate the use of these bundles in their characteristic registers.

Figure 5 provides a graphic representation of these data, showing the following overall patterns:

1 stance bundles are extremely common in both classroom teaching and conversation;
2 discourse organizing bundles are most common in classroom teaching and somewhat less common in conversation;
3 referential bundles are extremely common in classroom teaching and somewhat less common in textbooks and academic prose.

The patterns of use in classroom teaching are especially interesting here, and they help to explain why lexical bundles are generally so much more common in this register than any other register.[8] Classroom teaching

*Table 4: Distribution of common lexical bundles across functions*

| | Conversation | Classroom teaching | Textbooks | Academic prose |
|---|---|---|---|---|
| Stance bundles | 29 | 33 | 4 | 3 |
| Discourse organizers | 10 | 19 | 3 | 1 |
| Referential bundles | 3 | 32 | 20 | 15 |

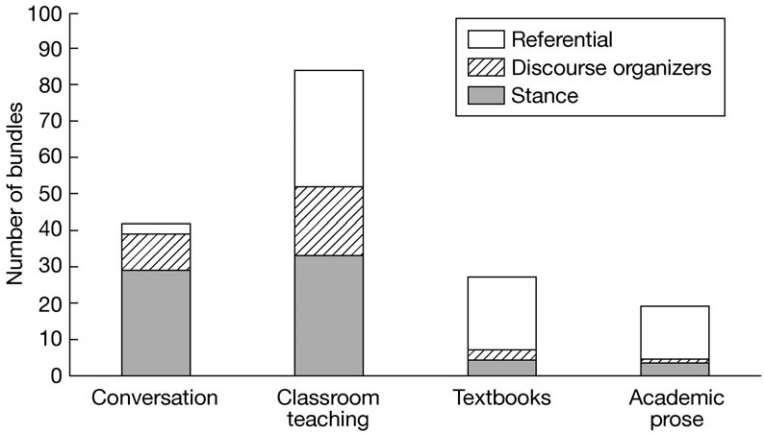


*Figure 5: Distribution of lexical bundles across functional categories*

combines functional characteristics of both conversation (using stance and discourse organizing bundles) and textbooks/academic prose (using referential bundles). However, classroom teaching actually goes beyond these other registers, using bundles in all three functional categories to a greater extent than any other register.

Two major patterns are noteworthy here: first, classroom teaching combines the functional and communicative priorities of involved spoken discourse (shown by the dense use of stance bundles) with the priorities of informational written discourse (shown by the dense use of referential bundles). Second, classroom teaching is structured with lexical bundles to a greater extent than any other register. This is shown most clearly by the large number of discourse organizing bundles used in classroom teaching. (Many of the referential bundles in classroom teaching are also used for discourse functions, such as identification/focus, imprecision, and quantity specification.) In fact, classroom teaching actually uses the most bundles in each functional category. This pattern apparently reflects the complex communicative demands of this register. Lexical bundles are useful for instructors who need to organize and structure discourse that is at once informational and involved, and is produced with real-time production constraints.

## 5.5 The relationship between structural and functional categories

Figure 6 shows that there is a very strong relationship between structural type and discourse function for lexical bundles. In particular, most stance bundles are composed of dependent clause fragments, while most referential bundles are composed of noun phrase or prepositional phrase fragments. (The prediction/intention stance bundles are all composed of VP fragments,

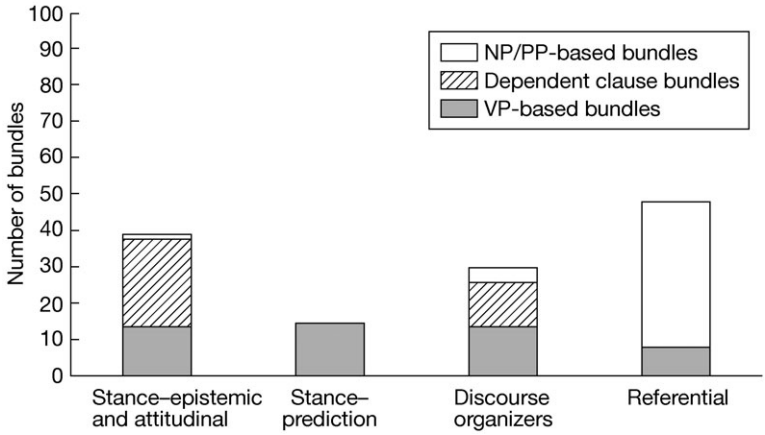


*Figure 6: Interaction of structural and functional categories*

mostly incorporating the semi-modal *be going to*.) Discourse organizers are the only functional category to use all three structural types.

Obviously, these patterns are strongly associated with register: conversation, at the 'oral' extreme, uses mostly VP-based and dependent clause lexical bundles for stance functions; academic prose, at the 'literate' extreme, uses mostly NP/PP-based lexical bundles for referential functions. Although classroom teaching is generally similar to conversation in its grammatical characteristics (see section 3 above), it uses NP/PP-based lexical bundles for referential purposes.

These patterns suggest that there is a direct association between form and function for lexical bundles. For example, complement clause-based bundles are used for stance functions, because complement clauses are generally one of the most productive grammatical devices used to express stance in English (see Biber *et al*. 1999, ch. 12). Thus, it makes sense that many of the most common multi-word sequences with complement clauses would become fixed as stance lexical bundles. Similarly, noun phrases and prepositional phrases are the primary grammatical devices used for referential functions, and so it makes sense that the most common of these multi-word sequences would become fixed as referential lexical bundles. Thus, we see a complex interaction between structural form, discourse function, and the typical purposes and situational characteristics of registers.

## 6 CONCLUSION: THE THEORETICAL STATUS OF LEXICAL BUNDLES

The results of the present analysis suggest that lexical bundles should be regarded as a basic linguistic construct with important functions for the construction of discourse. However, with respect to both structure and

function, lexical bundles differ dramatically from other linguistic features (including the traditional formulaic expressions usually recognized by discourse analysts).

Given that lexical bundles are defined strictly on the basis of frequency, with no consideration of structural or functional criteria, they might be expected to be arbitrary strings of words that have no linguistic status. Instead, these frequent sequences of words turn out to be readily interpretable in both structural and functional terms. Although they are not the kinds of grammatical structures recognized by traditional linguistic theory, most lexical bundles do have well-defined structural correlates: they usually consist of the beginning of a clause or phrase plus the first word of an embedded structure (e.g. a dependent complement clause or a prepositional phrase). These sequences of words can be regarded as structural 'frames', followed by a 'slot'. The frame functions as a kind of discourse anchor for the 'new' information in the slot, telling the listener/reader how to interpret that information with respect to stance, discourse organization, or referential status.

The patterns of use for lexical bundles are notably different from those found for traditional grammatical features. The contrast is especially striking for classroom teaching. With respect to grammatical features, classroom teaching relies heavily on 'oral' structures, despite the need for an informational focus. In our 2002 Multi-Dimensional study, we note that:

> On the spoken end, all university registers had scores indicating high involvement, reflecting their frequent use of such features as present tense verbs, private verbs, first- and second-person pronouns, and contractions. The most surprising inclusion in this group was classroom teaching, which had a notably involved rather than informational characterization. (Biber *et al.* 2002: 27)

These findings indicate that the typical grammatical characteristics of classroom teaching are determined primarily by the 'oral' characteristics of the situation: the real-time production circumstances, and the focus on personal and interpersonal purposes.

In contrast, in the present study we found that classroom teaching mixed 'oral' and 'literate' characteristics in the use of lexical bundles, actually going beyond the expected 'targets' in its patterns of use. That is, classroom teaching shows a more extensive use of stance lexical bundles and discourse organizing bundles than in conversation, while at the same time it shows a more extensive use of referential bundles than in academic prose.

These patterns of use indicate that lexical bundles are a fundamentally different kind of linguistic construct from productive grammatical constructions. For example, consider the use of NP/PP-based referential bundles in contrast to other noun- and prepositional-phrase structures. Classroom teaching relies heavily on NP/PP-based referential bundles but avoids the dense use of noun phrase and prepositional phrase structures in general. As a productive grammatical strategy, the dense use of complex noun phrase

constructions has dramatically increased in informational written registers over the past 100 years (see Biber and Clark 2002; Biber 2003). However, these constructions are much less common in spoken registers like conversation, presumably because they are difficult to produce and comprehend in real-time situations. In this regard, classroom teaching is typical of other 'oral' registers in avoiding the dense use of complex noun phrase structures.

Given those grammatical patterns, it is surprising that classroom teaching makes extensive use of NP/PP-based referential bundles. Although the functional need for referential expressions in classroom teaching is clear, the reliance on NP/PP-based bundles is unexpected. We interpret this pattern as evidence that lexical bundles are stored as unanalysed multi-word chunks, rather than as productive grammatical constructions. The fact that referential bundles are composed of noun phrase and prepositional phrase fragments reflects their historical origins, but we would argue that these sequences are now stored and used as single units, without reference to their structural correlates. As such, these bundles do not present production or comprehension difficulties for speakers and listeners in classroom teaching.

More general evidence for the importance of lexical bundles comes from their frequencies of use and obvious discourse functions. We have approached the study of lexical bundles, like all our corpus studies, with the general hypothesis that high frequency patterns are not accidental, nor are they explanatory. Rather, corpus-based frequency evidence provides descriptive facts that require explanation. In the present study, the facts that require explanation are the existence of common multi-word sequences which do not represent well-defined linguistic structures. Examination of these multi-word sequences in textual contexts shows that they are important building blocks of discourse, associated with basic communicative functions. In general, these lexical bundles serve as discourse framing devices: they provide a kind of frame expressing stance, discourse organization, or referential status, associated with a slot for the expression of new information relative to that frame.

Obviously, other approaches with different goals are important to complement the present study and increase our understanding of multi-word units, including psycholinguistic studies and investigations of more idiomatic and perceptually salient expressions.[9] At the same time, this study illustrates how an exploratory corpus approach facilitates the identification of language features that would go unrecognized otherwise, but that turn out to be a fundamentally important part of writers' and speakers' communicative repertoire.

*(Final version received April 2004)*

## NOTES

1 This study was carried out as part of the TOEFL 2000 Spoken and Written Academic Language Corpus project, supported by Educational Testing Service, with the goal of providing a basis for test construction and validation (see Biber *et al.* 2004).

2 The quantitative analysis of lexical bundles was undertaken with computer programs that identified and stored every four-word sequence in our corpus. The programs read through each text in the corpus, storing every sequence beginning with the first word of the text and advancing one word at a time. For example, the first sentence of this note would have the following four-word sequences identified: *the quantitative analysis of*, *quantitative analysis of lexical*, *analysis of lexical bundles*, *of lexical bundles was*, etc. Each time a sequence was identified, it was automatically checked against the previously identified sequences, and a running frequency count showed how often each sequence was repeated. In identifying lexical bundles, we relied on orthographic word units, even though these sometimes arbitrarily combine separate words. For example, *into*, *cannot*, *self-control*, and *don't* are all regarded as single words in our analysis. Only uninterrupted sequences of words were treated as lexical bundles. Thus, lexical sequences that spanned a turn boundary or a punctuation mark were excluded.

3 Other sequences of words can be repeated frequently within a single text. In many cases, these other sequences do not represent lexical bundles, because they are not widely used across multiple texts. These local repetitions reflect the immediate topical concerns of the discourse. In contrast, lexical bundles can be regarded as the more general lexical building blocks that are used frequently by many different speakers/writers within a register.

4 However, Simpson and Mendis (2003) document the important pragmatic functions of idioms in classroom teaching. These are generally rare, and they are often short noun phrases or prepositional phrases. For example, the most common idioms identified in this study (*bottom line* and *the big picture*) occur only 10 times per million words.

5 The quantitative findings for conversation and academic prose are taken from Biber *et al.* 1999; those for classroom teaching and textbooks are taken from Biber *et al.* (2004).

6 Figure 2 shows that textbooks uses a greater range of different lexical bundles than academic prose. Some of these bundles in academic prose occur with high frequencies, resulting in the slightly higher overall frequency of lexical bundles shown in Figure 3.

7 In fact, some of these bundles typically have stance functions in conversation, while they usually serve discourse organizing functions in classroom teaching. The bundle *I would like to* is a good example of this type.

8 Some differences in the sets of lexical bundles found across registers might be due to differences in the corpora analysed: ranging from *c.*5-million words for academic prose, to 1.2 million words for classroom teaching, and .75 million words for textbooks. However, Cortes (2002: 72–5) found that analyses of smaller corpora actually yield more lexical bundles, because some bundles have artificially high frequencies in the smaller corpora that cannot be maintained in a larger collection of texts. Thus, the low number of lexical bundles found in textbooks cannot be attributed to the fact that that sub-corpus is smaller; if anything, we would expect to find an artificially inflated number of bundles based on analysis of a smaller sub-corpus.

9 Schmitt *et al.* (2004) directly investigate the psycholinguistic status of lexical bundles through an experimental design. Their study indicates that not all bundles are stored in the mind as formulaic sequences, although register factors are not considered in the study. One longer-term goal is to extend the methods used to identify lexical bundles to allow for variations on a pattern. We considered the use of longer bundles and variations on a lexical pattern in Biber *et al.* (1999, ch. 13). In addition, speakers often use a series of expressions that are related in form and function, although they are not

strictly recurrent lexical sequences. For example, the following text sample illustrates variations on a bundle that incorporates the elements 1st/2nd person pronoun + *look at*:

> *If we look at* the fossils we find some very simple animals and *if you look at* more and more recent fo- fossils we find more and more complex animals and what I would like you to understand is what does this mean to increase in complexity *so we'll look at* what we call evolutionary trends or increases [writing on board] in complexity. and I'm gonna show you some slides of this in just a moment but basically *we can look at* animals from very simple animals to complex animals . . .

In our future research, we hope to interpret the use of lexical bundles relative to this wider range of lexical expressions. This type of study is much more feasible when focusing on a single lexical bundle in a few selected texts. The problem comes in trying to identify the full range of lexical bundles across a large corpus of texts.

## REFERENCES


**Aijmer, K.** 1996. *Conversational routines in English: Convention and creativity*. London: Longman.

**Akinnaso, F.** 1982. 'On the differences between spoken and written language.' *Language and Speech* 25: 97–125.

**Altenberg, B.** 1998. 'On the phraseology of spoken English: The evidence of recurrent word-combinations' in A. Cowie (ed.): *Phraseology: Theory, analysis and applications*. Oxford: Oxford University Press, pp. 101–22.

**Altenberg, B.** and **Eeg-Olofsson.** 1990. 'Phraseology in spoken English' in J. Aarts and W. Meijs (eds): *Theory and Practice in Corpus Linguistics*. Amsterdam: Rodopi, pp. 1–26.

**Biber, D.** 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

**Biber, D.** 1995. *Dimensions of Register Variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

**Biber, D.** 2003. 'Compressed noun phrase structures in newspaper discourse: The competing demands of popularization vs. economy' in J. Aitchison and D. Lewis (eds): *New Media Discourse*. London: Routledge, pp. 169–81.

**Biber, D.** and **V. Clark.** 2002. 'Historical shifts in modification patterns with complex noun phrase structures: How long can you go without a verb?' in T. Fanego, M. J. Lopez-Couso, and J. Perez-Guerra (eds): *English Historical Syntax and Morphology*. Amsterdam: John Benjamins, pp. 43–66.

**Biber, D.** and **S. Conrad.** 1999. 'Lexical bundles in conversation and academic prose' in H. Hasselgard and S. Oksefjell (eds): *Out of Corpora: Studies in honor of Stig Johansson.* Amsterdam: Rodopi, pp. 181–9.

**Biber, D., S. Conrad,** and **V. Cortes.** 2003. 'Lexical bundles in speech and writing: An initial taxonomy' in A. Wilson, P. Rayson, and T. McEnery (eds): *Corpus Linguistics by the Lune*. Frankfurt/Main: Peter Lang, pp. 71–93.

**Biber, D., S. Johansson, G. Leech, S. Conrad,** and **E. Finegan.** 1999. *The Longman Grammar of Spoken and Written English*. London: Longman.

**Biber, D., S. Conrad, R. Reppen, P. Byrd,** and **M. Helt.** 2002. 'Speaking and writing in the university: A multi-dimensional comparison.' *TESOL Quarterly*, 36, 9–48.

**Biber, D., S. Conrad, R. Reppen, P. Byrd, M. Helt, V. Clark,V. Cortes, E. Csomay,** and **A. Urzua.** 2004. *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus*. TOEFL Monograph Series. Princeton, NJ: Educational Testing Service.

**Brown, P.** and **C. Fraser.** 1979. 'Speech as a marker of situation' in K. R. Scherer and H. Giles (eds): *Social Markers in Speech.* Cambridge: Cambridge University Press, pp. 33–62.

**Butler, C. S.** 1997. 'Repeated word combinations in spoken and written text: Some implications for Functional Grammar' in C. S. Butler, J. H. Connolly, R. A. Gatward, and R. M. Vismans (eds): *A Fund of Ideas: Recent developments in Functional Grammar.*

Amsterdam: IFOTT, University of Amsterdam, pp. 60–77.

**Butler, C. S.** 1998. 'Collocational frameworks in Spanish.' *International Journal of Corpus Linguistics*, 3, 1–32.

**Chafe, W. L.** 1982. 'Integration and involvement in speaking, writing, and oral literature' in D. Tannen (ed.): *Spoken and Written Language: Exploring orality and literacy*. Norwood, NJ: Ablex, pp. 35–54.

**Chaudron, C.** and **J. Richards.** 1986. 'The effect of discourse markers on the comprehension of lectures.' *Applied Linguistics*, 7, 113–27.

**Chih-Hua, K.** 1999. 'The use of personal pronouns: Role relationships in scientific journal articles.' *English for Specific Purposes*, 18, 121–38.

**Conrad, S.** and **D. Biber.** In press. 'The frequency and use of lexical bundles in conversation and academic prose, in W. Teubert and M. Mahlberg (eds), 'The corpus approach to lexicography, Thematischer Teil von Lexicographica'. *Internationales Jahrbuch für Lexicographie* 20.

**Cortes, V.** 2002. Lexical bundles in academic writing in history and biology. Unpublished doctoral dissertation, Northern Arizona University.

**Cortes, V.** To appear. 'Lexical bundles in published and student disciplinary writing: Examples from history and biology.' *English for Specific Purposes*.

**Crompton, P.** 1997. 'Hedging in academic writing: Some theoretical problems.' *English for Specific Purposes*, 16, 271–87.

**DeCarrico, J.** and **J. Nattinger.** 1988. 'Lexical phrases for the comprehension of academic lectures.' *English for Specific Purposes*, 7, 91–102.

**deCock, 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.** 1996. 'Sequencing in SLA: Phonological memory, chunking, and points of order.' *Studies in Second Language Acquisition*, 19, 91–126.

**Flowerdew, J.** and **S. Tauroza.** 1995. 'The effect of discourse markers of second language lecture comprehension.' *Studies in Second Language Acquisition*, 17, 435–58.

**Francis, G., S. Hunston,** and **E. Manning.** 1996. *Collins COBUILD Grammar Patterns 1: Verbs*. London: HarperCollins.

**Francis, G., S. Hunston,** and **E. Manning.** 1998. *Collins COBUILD Grammar Patterns 2: Nouns and Adjectives*. London: HarperCollins.

**Francis, G., E. Manning,** and **S. Hunston.** 1997. *Verbs: Patterns and Practice*. London: HarperCollins.

**Gledhill, C.** 2000. 'The discourse function of collocation in research article introductions.' *English for Specific Purposes*, 19, 115–35.

**Grabe, W.** and **R. B. Kaplan.** 1996. *Theory and Practice of Writing*. London: Longman.

**Grabe, W.** and **R. B. Kaplan.** 1997. 'On the writing of science and the science of writing: Hedging in scientific text and elsewhere' in R. Markkanen and H. Schroeder (eds): *Hedging in Discourse*. Berlin: de Gruyter, pp. 151–67.

**Granger, S.** 1998. 'Prefabricated patterns in advanced EFL writing: collocations and formulae' in A. Cowie (ed.): *Phraseology*. Oxford: Oxford University Press, pp. 145–60.

**Gumperz, J. J., H. Kalman,** and **M. C. O'Conner.** 1984. 'Cohesion in spoken and written discourse' in D. Tannen (ed.): *Coherence in Spoken and Written Discourse*. Norwood, NJ: Ablex, pp. 3–20.

**Hakuta, K.** 1974. 'Prefabricated patterns and the emergence of structure in second language acquisition.' *Language Learning*, 24, 287–97.

**Halliday, M. A. K.** 1978. *Language as Social Semiotic: The social interpretation of language and meaning*. London: Edward Arnold.

**Halliday, M. A. K.** 1988. 'On the language of physical science' in M. Ghadessy (ed.): *Registers of Written English*. London: Pinter, pp. 162–78.

**Halliday, M. A. K.** and **J. R. Martin.** 1993. *Writing Science: Literacy and discursive power*. Pittsburgh: University of Pittsburgh Press.

**Howarth, P.** 1996. *Phraseology in English Academic Writing*. Tübingen: Max Niemeyer Verlag.

**Howarth, P.** 1998a. 'Phraseology and second language proficiency.' *Applied Linguistics*, 19, 24–44.

**Howarth, P.** 1998b.'The phraseology of learners' academic writing' in A. Cowie (ed.): *Phraseology*. Oxford: Claredon Press, pp. 161–86.

**Hudson, J.** 1998. *Perspectives on Fixedness:*

*Applied and theoretical*. Lund: Lund University Press.

**Hunston, S.** and **G. Francis.** 1998. 'Verbs observed: a corpus-driven pedagogic grammar.' *Applied Linguistics*, 19, 45–72.

**Hunston, S.** and **G. Francis.** 1999. *Pattern Grammar: A Corpus-driven approach to the lexical grammar of English*. Amsterdam: Benjamins.

**Hyland, K.** 1994. 'Hedging in academic writing and EAP textbooks.' *English for Specific Purposes*, 13, 239–56.

**Hyland, K.** 1996a. 'Talking to the academy: Forms of hedging in science research articles.' *Written Communication*, 13, 251–81.

**Hyland, K.** 1996b. 'Writing without conviction? Hedging in science research articles.' *Applied Linguistics*, 17, 433–54.

**Hymes, D.** 1974. *Foundations in Sociolinguistics*. Philadelphia: University of Pennsylvania Press.

**Kjellmer, G.** 1991. 'A mint of phrases' in K. Aijmer and B. Altenberg (eds): *English Corpus Linguistics: Studies in honour of Jan Svartvik*. London: Longman, pp. 111–27.

**Kroll, B.** 1977. 'Ways communicators encode propositions in spoken and written English: A look at subordination and coordination' in E. O. Keenan and T. Bennett (eds): *Discourse across Time and Space*. Los Angeles: University of Southern California, pp. 69–108.

**Lewis, M.** 1993. *The Lexical Approach: The state of ELT and a way forward*. Hove: LTP.

**Marco, M. J. L.** 2000. 'Collocational frameworks in medical research papers: a genre-based study.' *English for Specific Purposes*, 19, 63–86.

**Moon, R.** 1998a. *Fixed Expressions and Idioms in English: A corpus-based approach*. Oxford: Clarendon.

**Moon, R.** 1998b. 'Frequencies and forms of phrasal lexemes in English' in A. Cowie (ed.): *Phraseology*. Oxford: Oxford University Press, pp. 79–100.

**Myers, G.** 1989. 'The pragmatics of politeness in scientific articles.' *Applied Linguistics*, 10, 1–35.

**Nattinger, J.** and **J. DeCarrico.** 1992. *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

**O'Donnell, R. C.** 1974. 'Syntactic differences between speech and writing.' *American Speech* 49, 102–10.

**Partington, A.** 1998. *Patterns and Meanings: Using corpora for English language research and teaching*. Amsterdam: Benjamins.

**Partington, A.** and **J. Morley.** 2004. 'From frequency to ideology: Investigating word and cluster/bundle frequency in political debate' in B. Lewandowska-Tomaszczyk (ed.): *Practical Applications in Language and Computers—PALC 2003*. Frankfurt a. Main: Peter Lang, pp. 179–92.

**Pawley, A.** and **H. Syder.** 1983. 'Two puzzles for linguistic theory: Native-like selection and native-like fluency' in J. Richards and R. Schmidt (eds): *Language and Communication*. London: Longman, pp. 191–226.

**Redeker, G.** 1991. 'Review article: Linguistic markers of discourse structure.' *Linguistics*, 29, 1139–72.

**Renouf, A.** and **J. Sinclair.** 1991. 'Collocational frameworks in English' in K. Aijmer and B. Altenberg (eds): *English Corpus Linguistics: Studies in honour of Jan Svartvik*. London: Longman, pp. 128–43.

**Salem, A.** 1987. *Pratique des Segments Répétés*. Paris: Institut National de la Langue Française.

**Schmitt, N.** (ed.). 2004. *Formulaic Sequences*. Amsterdam: John Benjamins.

**Schmitt, N., S. Grandage** and **S. Adolphs.** 2004. 'Are corpus-derived recurrent clusters psycholinguistically valid?' In N. Schmitt (ed.), *Formulaic Sequences*. Amsterdam: John Benjamins. pp. 127–51.

**Simpson, R.** and **D. Mendis.** 2003. 'A corpus-based study of idioms in academic speech.' *TESOL Quarterly*, 37, 419–41.

**Sinclair, J.** 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

**Strodt-Lopez, B.** 1991. 'Tying it all in: Asides in university lectures.' *Applied Linguistics*, 12, 117–40.

**Swales, J. M.** 1990. *Genre Analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

**Swales, J. M., U. K. Ahmad, Y. Y. Chang, D. Chavez, D. F. Dressen,** and **R. Seymour.** 1998. 'Consider this: The role of imperatives in scholarly writing.' *Applied Linguistics*, 19, 97–121.

**Thompson, G.** and **Y. Ye.** 1991. 'Evaluation in the reporting verbs used in academic papers.' *Applied Linguistics*, 12, 365–82.

**Weinert, R.** 1995. 'The role of formulaic language in second language acquisition: A review.' *Applied Linguistics*, 16, 180–205.

**Williams, I.** 1996. 'A contextual study of lexical verbs in two types of medical research report: Clinical and Experimental.' *English for Specific Purposes*, 15, 175–97.

**Wray, A.** and **M. Perkins.** 2000. 'The functions of formulaic language: an integrated model.' *Language and Communication*, 20, 1–28.

**Wray, A.** 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

**Yorio, C.** 1980. 'Conventionalized language forms and the development of communicative competence.' *TESOL Quarterly*, 14, 433–42.