Susan Conrad / Douglas Biber

The Frequency and Use of Lexical Bundles in Conversation and Academic Prose

1 Introduction

1.1 The importance of multi-word sequences for lexicography

Even before the use of computer-assisted techniques in lexicography and linguistics, scholars interested in language use recognized the importance of recurring patterns. FIRTH (1957, 195) noted that patterns in the surrounding context were important for understanding the meaning of a word, stating "you shall know a word by the company it keeps". In looking at the social functions of language, HYMES (1968, 126) claimed that "a vast portion of verbal behavior … consists of recurrent patterns, of linguistic routines". Branches of lexicology, too, for decades have investigated the status of multi-word units (see review in MOON 1997, 48–50). Nevertheless, lexicography continues to emphasize the individual word as the basic unit of discourse. The very fact that dictionaries are arranged by individual head words gives primacy to the individual word, and suggests that phrases and clauses of a language are built from these individual units. Particularly in the United States, with the strong influence of CHOMSKYan linguistics and its emphasis on syntactic rules for generating all utterances, multi-word sequences have received little emphasis.

As empirical work with multi-word sequences has increased, however, it has become impossible to ignore their importance for describing the lexicon of a language. First, as many scholars have pointed out (e.g. PAWLEY/SYDER 1983; SINCLAIR 1991; WRAY/PERKINS 2000), if individual words were indeed the building blocks of language—combined through the application of syntactic rules—we should see a great deal of novel language use, with innovative phrases and clauses. Instead, much language use consists of repeated expressions. This fact has become particularly obvious as corpus-based research has been used in lexical studies. Depending on the definition given to formulaic language use (a

Brought to you by | Portland State University (Portland State University)
Authenticated | 172.16.1.226
Download Date | 3/27/12 1:02 AM

matter discussed further below), estimates have been as high as 80% of the words in a corpus being within recurrent sequences (ALTENBERG 1998). Working with evidence from corpora has led SINCLAIR (1991, 110) to posit the "idiom principle": that even though phrases may appear to be analyzable into smaller segments, "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices".

The idea that humans store multi-word sequences as single units seems reasonable not only from a frequency-of-use perspective, but also from a psycholinguistic one. Retrieving and recognizing such multi-word units would facilitate the level of fluency that speakers exhibit even with processing pressures, such as time constraints or attention given to other tasks. Empirical studies within psycholinguistics support this notion as well. WRAY (2002, chapters 1 and 2) reviews work on phonological patterns, including stress and the timing of pauses, and the placements of code switching, which lend support to the idea that human language production and processing includes multi-word sequences as single units. WRAY (2002) posits a dual system for language processing, arguing that while we do have the capacity for analytical processing, our preferred mode of processing is more holistic because it requires less processing effort. While more direct evidence is needed, it is clear that, for psycholinguists too, multi-word sequences are an important consideration for language descriptions.

1.2 Identifying multi-word sequences

Even scholars who agree on the importance of multi-word sequences often disagree about the methods for identifying and studying them. Six characteristics of multi-word sequences tend to be singled out as especially important: fixedness; idiomaticity; frequency; length of sequence; completeness in syntax, semantics, or pragmatics; and intuitive recognition by native speakers of a language community. However, studies give differential priority to these characteristics, depending on the focus of the study.

For example, a study of idioms such as *kick the bucket* (meaning "die") or *a slap in the face* (meaning "an affront") gives priority to fixedness, idiomaticity, completeness, and intuitive recognition by native speakers—i.e., idioms tend to have a fixed form (thus we do not say *kicked a bucket*), the meaning is not transparent from the parts, they tend to be complete phrases and are semantically complete (though they can be any length), and native speakers recognize the sequence as a unit. However, idioms are quite rare in normal use; *kick the bucket* and *a slap in the face*, for example, are rarely attested in typical face-to-face conversation (see BIBER/JOHANSSON/LEECH/CONRAD/FINEGAN 1999, 1024–6). Studies focused on idioms, therefore, miss a majority of recurring multi-word sequences.

In contrast to idioms, studies of collocations give primacy to frequency and two-word relationships. For example, KENNEDY (2003) analyzes combinations of degree adverbs + adjectives in the British National Corpus (e.g. *extremely rare, greatly appreciated*). High frequency adverbs of degree are chosen as the focus, and the strength of their collocations is analyzed statistically, using a two-word window on each side so that intervening words are allowed. The sequences therefore do not have to be fixed in form. Whether they are intuitively recognized by native speakers is unimportant, since the combinations are identified statistically. Idiomaticity is not a criterion since the meaning is discernable from the parts, but the collocations do form a complete semantic unit.

Brought to you by | Portland State University (Portland State University)
Authenticated | 172.16.1.226
Download Date | 3/27/12 1:02 AM

Studies of collocations have also been expanded to non-contiguous sequences, having a fixed lexical frame combined with an open slot. For example, RENOUF/SINCLAIR (1991) describe "collocational frameworks"—pairs of words separated by one intervening word, e.g. *a* + ___ + *of*. It turns out that such frames tend to collocate with semantically similar words (for example, *a* + ___ + *of* is associated with words of measurements or parts), but the sequences under study clearly are not syntactically complete. (See also BUTLER 1998, on collocational frameworks in Spanish.)

Other studies give primacy to intuitive recognition coupled with semantic or pragmatic completeness. NATTINGER/DECARRICO (1992) focus on "lexical phrases," which they identify as fixed expressions used for a clear pragmatic function. They give examples of types of lexical phrases that are useful for English as a second language learners—e.g. the "sentence builder" *there is no doubt that* or "topic marker" *what I mainly wanted to talk to you about was*—without attempting quantitative or exhaustive lists. While clearly useful for pedagogical purposes, the drawback of such an approach for lexicographic description is that we do not know how many such sequences occur in natural discourse without our having consciously noticed them.

Given the variety of purposes in studies of multi-word sequences, it is not surprising that the sequences have been identified in diverse ways (including many more than have been described here—e.g. MOON (1998a and b) studies sequences identified as "phrases" in previous lexicographic work, and ERMAN/WARREN (2000) identify prefabricated sequences with a combined emphasis on semantic criteria and intuitive judgments). Ultimately, the criteria used to identify multi-word sequences are tied to the purpose of the study. Describing associations between words requires identifying different types of sequences than listing the most pedagogically useful sequences does, and both of these differ from cataloging fixed idioms in a language.

In our previous work with multi-word sequences—which we have labeled a "lexical bundle" approach (BIBER et al. 1999; BIBER/CONRAD 1999; BIBER/CONRAD/CORTES 2003; BIBER/CONRAD/CORTES 2004)—our purpose has been to identify the most common recurrent sequences of words and to determine the extent to which those sequences can be interpreted as building blocks of discourse. Our research questions in this approach are exploratory. We ask whether there are multi-word sequences that are used with high frequency in texts, whether different registers tend to use different sets of these sequences, and, if so, to what extent the bundles fulfill discourse functions and thus play an important part in the communicative repertoire of speakers and writers.

The purposes of the lexical bundle approach require that multi-word sequences be identified with priority given to frequency, fixedness, and sequences longer than two words. Our hypothesis is that extremely common, fixed sequences of words are used as unanalyzed chunks by speakers and writers, and therefore will have identifiable discourse functions in texts. Sequences of two words are not included since many of them are word associations that do not have a distinct discourse-level function. In many cases, lexical bundles are not structurally complete, and they are not units that linguists would recognize using their intuition. For example, *I don't know if*, *it is possible to*, and *the nature of the* are all common recurrent sequences of words, but they are unlikely to be recognized as complete lexical chunks based on intuition. Nevertheless, as we show in the following sections, it turns out that the lexical bundles identified purely on frequency criteria do have strong functional

Brought to you by | Portland State University (Portland State University)
Authenticated | 172.16.1.226
Download Date | 3/27/12 1:02 AM

correlates, indicating that speakers and writers regularly use them as basic building blocks of discourse.

In using a frequency-driven, fixed-word approach to identifying multi-word sequences, we follow ALTENBERG (1993, 1998), who carried out initial work of this type with spoken texts in the London-Lund corpus. BUTLER (1997) applies this approach to Spanish. In the *Longman Grammar of Spoken and Written English* (BIBER et al. 1999, chapter 13; hereafter the *Longman Grammar*), we emphasized the structures of lexical bundles, and discussed the structures' associations with various discourse functions. Here we summarize major findings of that work and then extend it, presenting an initial classification of the lexical bundles into functional categories. We continue to adopt a register perspective—comparing the bundles across different varieties of language based on their contexts of use. In the present paper, we focus on the comparison of conversation and academic prose in English, though the methodology can be applied to other registers as well as other languages.

2 Methodology

2.1 Corpus used for the study

The present paper summarizes and extends findings about lexical bundles analyzed in the Longman Spoken and Written English Corpus (Table 1). The analysis summarized here considers only the British English component of the conversation subcorpus. Sampling was carried out along demographic guidelines to represent a range of speakers in the UK, with participants recording all of their conversations over a week. Approximately 500 speakers are included. The academic prose subcorpus includes research articles and book extracts; most book extracts come from trade books written for an audience with some technical background, but about 20% come from books written for a lay audience, including student textbooks. More details about the corpus can be found in the *Longman Grammar* (chapter 1).

Register	Number of Texts	Number of Words
Conversation (British English)	3,436	3,929,500
Academic Prose (American and British English)	408	5,331,800

Tab. 1: Components of the Longman Spoken and Written English Corpus used in the present study

2.2 Identification and frequency counts of lexical bundles

We define lexical bundles as the most frequent recurring fixed lexical sequences in a register. The more common a lexical bundle, the more useful it would appear to be in building discourse, but precisely where to set a frequency cut-off is somewhat arbitrary. We give an

Brought to you by | Portland State University (Portland State University)
Authenticated | 172.16.1.226
Download Date | 3/27/12 1:02 AM

overall summary of the frequency of 3- and 4-word lexical bundles considering bundles with a frequency of at least 10 per million words in the register, but for the sake of brevity the discussion of the functional classifications focuses on bundles that occur at least 40 times per million words.

To be considered a lexical bundle, the sequence must also be used by multiple speakers or authors, and not simply be a matter of individual style. A lexical bundle must thus occur across numerous texts in a register. We used a cut-off of occurrence in at least 5 different texts for the analyses here, though this limit had little practical effect because most bundles are widely distributed; most of the bundles included here are found in more than 30 texts.

The frequency analysis for the lexical bundles was undertaken with computer programs that identified and stored every sequence in the corpus, with three-word, four-word, five-word, and six-word lists kept separately. Contractions were considered single words. Each text in the corpus was read through word by word, storing every sequence beginning with the first word of the text and advancing one word at a time. For example, to identify four-word lexical bundles, the beginning of this paragraph would be processed as

> the frequency analysis for
> frequency analysis for the
> analysis for the lexical

Each sequence was checked against the previously identified sequences, and a running count was kept of how often the sequence was repeated, along with the number of different texts it occurred in. Sequences that recurred above the cut-off for both overall frequency and the number of texts were included as lexical bundles.

It is important to remember that lexical bundles do not have to represent complete structural units or semantic units. Clause and phrase boundaries were not separated in the analysis, so a lexical bundle may bridge two syntactic units (as in fact happens with many bundles, such as *you know what I* in conversation). The analysis of the structural and functional characteristics of the bundles took place after their identification.

2.3 Classification of lexical bundles

The lexical bundles are classified in two major ways. First, we consider the structural characteristics of the bundles. Although most of the bundles are not complete structural units, they do fall into groups with certain structural associations. For example, bundles like *you want me to* are constructed from verb and clause components, while bundles like *in the case of* are constructed from noun phrase and prepositional phrase components. After the bundles were identified by the frequency analysis, they were fully categorized into 12 structural types described in the *Longman Grammar*, taking into account the initial elements of the bundle and its overall structure. For brevity here, we present only the structures that account for at least 10% of the 4-word lexical bundles in each register.

The second type of classification presented in this paper is a preliminary classification of the bundles by their function in a discourse context. No a priori categories were assumed. Instead, we examined each bundle in concordance listings and made interpretations of its function. We placed bundles into groups unified by similar discourse functions.

Brought to you by | Portland State University (Portland State University)
Authenticated | 172.16.1.226
Download Date | 3/27/12 1:02 AM

In some cases, the function of a lexical bundle is clear even out of context. For example, *it is necessary to* conveys an obligation in an impersonal way. For many bundles, however, only looking at their occurrence in context can determine the function. Some are multifunctional. For example, the bundle *the end of the* can function as a time reference or place reference. We have classified these bundles under their most common function, or when more than one function is common, have made multifunctional categories (such as multifunctional time/place reference).

Although the discussion of functions in this paper highlights only a few points about the most common four-word bundles, the system of classification was developed by analysis of all four-word bundles in four registers—conversation, academic prose, university textbooks, and university class sessions (see BIBER/CONRAD/CORTES 2004).

3 The frequency of lexical bundles in conversation and academic prose

The frequency of lexical bundles can be examined from several perspectives, most notably the diversity in lexical bundles, how frequently these bundles are used in discourse, and what percentage of words in the discourse are contained in lexical bundles. All three of these perspectives speak to the importance of lexical bundles as building blocks of discourse.

Taking three-word and four-word lexical bundles together, there are almost 4,000 different lexical bundles in conversation, and about 3,000 different lexical bundles in academic prose (Figure 1). Overall, these bundles occur very commonly in both registers. In conversation, three-word bundles occur over 80,000 times per million words and four-word bundles over 8,500 times per million words. In academic prose, three-word bundles occur over 60,000 times per million words, and four-word bundles over 5,000 times per million words. In frequency of use of individual bundles, however, the registers are quite different. Conversation has a few bundles with very high frequencies. For example, the three-word bundle *I don't know* occurs over 1,000 times per million words. On the other hand, the most common bundles in academic prose occur between 200 and 400 times per million words (e.g. *in order to*, *one of the, part of the*). In terms of the proportion of text covered by lexical bundles, conversation is also higher than academic prose: approximately 28% of the words in conversation occur within 3- and 4-word lexical bundles, while in academic prose the percentage is about 20% (Figures 2 and 3).

Brought to you by | Portland State University (Portland State University)
Authenticated | 172.16.1.226
Download Date | 3/27/12 1:02 AM

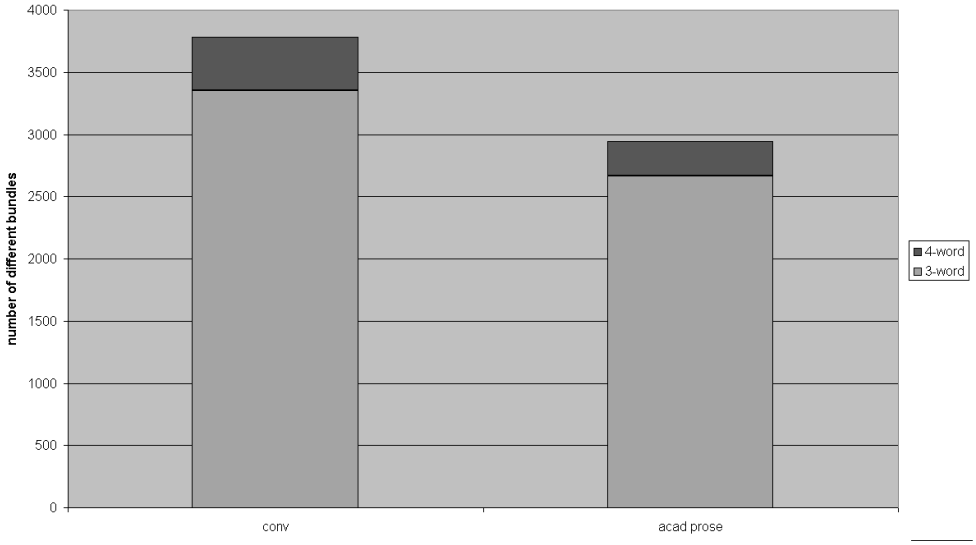

Fig. 1: Number of lexical bundles in conversation and academic prose

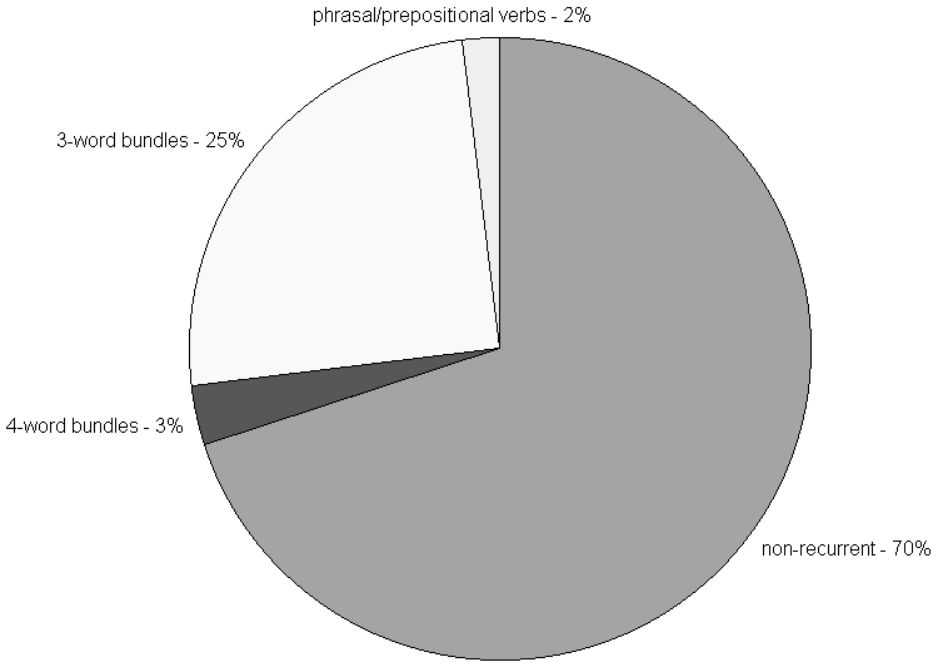

Fig. 2: Percentage of words in conversation in lexical bundles

Brought to you by | Portland State University (Portland State University)
Authenticated | 172.16.1.226
Download Date | 3/27/12 1:02 AM

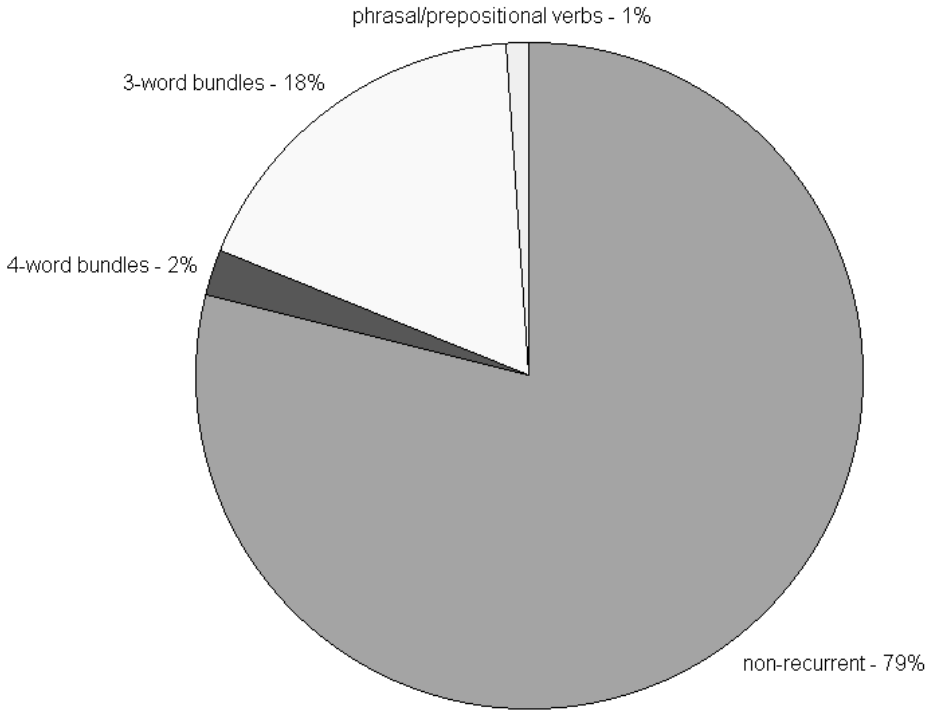

Fig. 3: Percentage of words in academic prose in lexical bundles

Of course, the exact frequencies are highly dependent upon the definition for lexical bundles, since different frequency cut-offs would result in a different number of bundles. Even with that caveat, however, the findings raise an important issue for descriptions of language use. With these bundles accounting for 1/5 to 1/4 of the occurrences of words, it seems unlikely that their recurrent use is a matter of chance. Although previous work has sometimes emphasized the importance of repeated expressions in speech (e.g., McCarthy/Carter 1997), this study shows that even in academic prose, these recurrent bundles appear to be useful building blocks for the discourse. The structural and functional differences across the registers further show that the bundles fulfill communicative purposes that are particularly important for each of the registers.

4 Structural patterns of the lexical bundles

Although most of the bundles are not complete syntactic structures, a major difference in their structure across the registers is clear. Bundles in conversation are most commonly parts of declarative clauses or questions; about 90% of the lexical bundles include part of a

Brought to you by | Portland State University (Portland State University)
Authenticated | 172.16.1.226
Download Date | 3/27/12 1:02 AM

verb phrase. In academic prose, on the other hand, more lexical bundles (about 60%) are parts of noun phrases and/or prepositional phrases.

Considering the structures that account for at least 10% of the 4-word bundles in each register illustrates the contrast between the registers. (The *Longman Grammar*, chapter 13, provides a complete review of the structures.) Three structural types account for almost 70% of the 4-word bundles in conversation (Table 2), and all three include a verb. However, these structures account for only a negligible proportion of the bundles in academic prose. Rather, over 60% of the 4-word bundles in academic prose are covered by two structural types that incorporate noun phrase components; these structures account for only about 7% of the bundles in conversation.

The difference in lexical bundle structures between the registers is consistent with word, phrase and clause category differences between these registers generally. Conversation tends to have more verbs, more personal pronouns, and more questions, while academic prose has more nouns and prepositional phrases (*Longman Grammar*, chapters 2, 8, 14). More importantly, these structural differences reflect differences in the functions that the bundles serve. The structures typical of conversation are used for more personal expressions, particularly expressions of attitudes and desires, with bundles such as *I don't know what* or *you want me to*. The structures typical of academic prose are useful for specifying aspects of information with bundles such as *the nature of the*, *the extent to which*, and *as a result of*. These functional differences provide greater insight into lexical bundles' role in building discourse, and are the focus of the next section.

Percent	Structural Type	Example
Conversation		
44%	personal pronoun + lexical VP (+ complement clause)	*I don't know what*
13%	(aux +) active V (+)	*have a look at*
12%	yes-no and wh-question fragments	*can I have a* *what do you want*
Academic Prose		
33%	preposition + NP fragment	*as a result of*
30%	NP with post-modifier fragment	*the nature of the*

Tab. 2: Structures accounting for at least 10% of the 4-word lexical bundles

5 Functional classification of the lexical bundles

The functional classification of the bundles resulted in four general categories: stance expressions, discourse organizers, referential expressions, and special conversational functions. Before discussing the differences across the registers in the next section, we describe

Brought to you by | Portland State University (Portland State University)
Authenticated | 172.16.1.226
Download Date | 3/27/12 1:02 AM

each category and its subcategories. Table 3 lists the common 4-word bundles in each of the categories (bundles occurring over 40 times per million words).

– *Stance Expressions*: Stance bundles express attitudes or assessments that provide a frame for the interpretation of the following proposition, such as *I don't know if* and *it is necessary to*. They convey 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 what*, *I don't think so, the fact that the*). Attitudinal/modality stance bundles express speaker attitudes towards the actions or events described in the following proposition (e.g. *I don't want to, I'm not going to*). We found four types of attitudinal/modality bundles—focused on desire (e.g. *I don't want to*), obligation/directive (e.g. *you don't have to, it is necessary to*), intention/prediction (e.g. *I was going to*, *it's going to be*), and ability (*it is possible to*). Stance bundles are also classified by whether they convey the stance in a personal or impersonal way. Personal stance bundles overtly attribute the stance to the speaker/writer or addressee (*you* or *I*). Impersonal stance bundles express similar meanings without being attributed directly to an individual (e.g. *it is possible to*).

	Conversation	**Academic Prose**
I. STANCE EXPRESSIONS		
I-A. Epistemic Stance		
Personal:	I don't know what I don't know if I don't know how I think it was 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:		the fact that the
I-B. Attitudinal/Modality Stance		
Desire	I don't want to do you want to if you want to you want to go do you want a I would like to what do you want	
Obligation/Directive		
Personal:	you don't have to you don't want to going to have to you want me to do you want me	
Impersonal:		it is necessary to
Intention/Prediction		

Brought to you by | Portland State University (Portland State University)
Authenticated | 172.16.1.226
Download Date | 3/27/12 1:02 AM

	Conversation	Academic Prose
Personal:	I was going to are you going to I'm not going to are we going to	
Impersonal:		it's going to be going to be a going to have a
Ability		
Impersonal:		it is possible to
II. DISCOURSE ORGANIZERS		
II-A. Topic Introduction/Focus	what do you think do you know what I/I'll tell you what have a look at let's have a look	
II-B. Topic Elaboration/Clarification	nothing to do with know what I mean was going to say what do you mean	on the other hand
III. REFERENTIAL EXPRESSIONS		
III-A. Identification/Focus		one of the most
III-B. Imprecision	or something like that	
III-C. Specification of Attributes		
Quantity Specification		per cent of the
Tangible Framing Attributes	in the form of	
Intangible Framing Attributes		in the case of the nature of the as a result of on the basis of in the absence of the way in which the extent to which in the presence of
III-D. Time/Place/Text Reference		
Time Reference		at the same time at the time of
Multi-Functional Reference	the end of the at the end of	the end of the at the end of
IV. SPECIAL CONVERSATIONAL FUNCTIONS		
Politeness	thank you very much	
Simple Inquiry	what are you doing	
Reporting	I said to him	

Tab. 3: Functional Classification of 4-word lexical bundles with frequencies over 40/million words

Brought to you by | Portland State University (Portland State University)
Authenticated | 172.16.1.226
Download Date | 3/27/12 1:02 AM

- *Discourse Organizers*: Discourse organizers reflect relationships between prior and coming discourse. They serve two major functions: topic introduction/focus and topic elaboration/clarification. Topic introduction/focus bundles provide overt signals that a new topic (or subtopic) is being introduced or is becoming the focus of attention (e.g. *do you know what, I tell you what*). Topic elaboration/clarification bundles serve to add more information to a topic (e.g. *nothing to do with*) or to clarify or ask for clarification of previously stated information (e.g. *what do you mean*). They can also overtly mark the relationship the speaker/writer sees between units of discourse, as with *on the other hand*.
- *Referential Expressions*: Referential bundles make direct reference to physical or abstract entities, or to the textual context. We found four types. Identification/focus bundles identify an entity or part of it as noteworthy (*one of the most*). Imprecision bundles communicate that previous discourse is expressed imprecisely (and they are thus related to stance expressions which convey uncertainty—e.g. *or something like that*). Bundles in the "specification of attributes" category bring focus to some particular attribute of the entity, including quantities (*per cent of the*), tangible attributes (*in the form of*), and intangible attributes (e.g. *the nature of the, in the absence of, the way in which*). Time/place/text references can refer to one of those areas or be multi-functional (e.g. *the end of the*).
- *Special Conversational Functions*: The special conversational functions cover three subcategories that occurred only in the conversation subcorpus: politeness routines (*thank you very much*), simple inquiry (*what are you doing*), and reporting clauses (*I said to him*).

Interestingly, as Table 3 shows, the common 4-word bundles in conversation and academic prose have almost entirely non-overlapping distributions with respect to the functional categories. Both registers have at least one elaboration/clarification bundle. Both also have two multi-functional reference bundles—and these are the only common 4-word bundles that both registers share (*the end of the* and *at the end of*).[1] Otherwise, the functions typically fulfilled by the common bundles in each register are quite distinct.

5.1 The function of common lexical bundles in conversation

The functional types of bundles that are common in conversation reflect the communicative purposes and contexts of typical conversation in British English—a focus on interaction and conveying personal thoughts and attitudes, and the concern for politeness and not imposing on others. The most striking aspect of conversation's use of lexical bundles is the high proportion of personal stance expressions. They are used for epistemic stance (usually expressing lack of certainty or knowledge); expressing personal desires and inquiring into others' desires; directing others, releasing them from obligations, or inquiring into one's own obligations; and discussing intentions. Examples include:

I don't know how you got on that list. [epistemic stance]
I don't want to go by myself. [attitude/modality—desire]
You sure *you want to go*? [attitude/modality—desire]
As soon as you've finished just go, *you don't have to* stay for your full three hours, nobody's gonna know [attitude/modality—obligation/directive]
A: She can't cope.
B: Oh dear. What *are we going to* do now then? [attitude/modality—intention/prediction]

1 In some cases, 4-word bundles are parts of 5-word or 6-word bundles (e.g. *at the end of* and *the end of the* are both part of *at the end of the*). These longer bundles are far less common and, for brevity, are not covered here (see further the *Longman Grammar*, chapter 13).

Brought to you by | Portland State University (Portland State University)
Authenticated | 172.16.1.226
Download Date | 3/27/12 1:02 AM

The usefulness of the stance bundles in conversation extends beyond simply conveying stance. The expressions of stance have important interactional functions, for example, using an epistemic stance bundle to reduce the imposition of one's own opinion:

> You know you can't be hard on people can you really? Or *I don't think so* anyhow…

They are also used in indirect questions, drawing others into the conversation and/or showing concern for shared background knowledge, for example:

> A: There was a programme on this morning, *I don't know if* you saw it [or not?]
> B: [No I didn't] no

The bundles in the discourse organizing category, though fewer, also reflect the interactive concerns of conversation. For example, some topic introducing/focusing bundles are question structures which draw the reader in and others are expressions that encourage joint action:

> A: So what you do, *what do you think* I should do when I see Mary tomorrow?
> B: Give er a right smack across the chops
> *Do you know what* I did over the weekend?
> You've bought two, well *let's have a look* at yours mate then

In addition, topic elaboration/clarification bundles mostly have to do with clarification, for example:

> A: Are you gonna have the house or not in London? Er, the flat?
> B: I might.
> A: *What do you mean*, might? We need to know.

5.2 The function of common lexical bundles in academic prose

The majority of the common four-word bundles in academic prose are referential expressions. The most common subcategory is the specification of attributes, with bundles covering quantities, tangible attributes and a variety of intangible attributes, for example:

> Approximately 80 *per cent of the* respondents claimed to possess some feelings of attachment to a "home" community area…
> When a monomer polymerizes it will only yield a useful polymer if it does so *in the form of* a long chain.
> Candidates are selected *on the basis of* technical qualifications.
> *The extent to which* individuals are able to participate in employment, leisure and social interaction, for example, will be an indicator of the reality of their "adulthood".

In contrast to conversation, academic prose has only three common four-word bundles that express stance, and all of them are impersonal. Furthermore, unlike conversational epistemic stance bundles, which usually convey uncertainty, the one epistemic stance bundle common in academic prose—*the fact that the*—focuses on certainty. Often this bundle encapsulates a concept and presents it as established, accepted information, as in the following:

Brought to you by | Portland State University (Portland State University)
Authenticated | 172.16.1.226
Download Date | 3/27/12 1:02 AM

> …the best basis for adaptability is a liberal education aimed at generating a wide understanding and the development of reason and autonomy. *The fact that the* more complex realms of human action and reflection are the most important and valuable should inhibit us in assuming that all education should issue in behaviourally identifiable skills.

It may be surprising that academic prose does not have more common bundles in the discourse organizing category. However, topic introductions and elaborations take place with a greater variety of expressions as well as with two-word expressions such as *for example*, rather than with lexical bundles. The one common discourse organizing bundle, which is used for explicit contrast, is one of the few structurally complete bundles—*on the other hand*.

The great difference between the functions of the common bundles in academic prose and conversation does not mean, of course, that academic prose does not convey stance or have certain conventions for politeness strategies. But the lexical bundles are used for the most important and overt considerations of the register. Academic prose puts a premium on the conveying of precise information, over the interpersonal considerations of a face-to-face interaction, and it is the focus on information that is apparent in the lexical bundles.

6 Conclusion

We characterized the lexical bundle approach as an exploratory approach to multi-word sequences, first asking whether there even are multi-word sequences used with a high frequency by speakers and writers. The answer to this question is clearly "yes." The fact that previous work has not identified many of the lexical bundle sequences might make one wonder if their repetition is just accidental. However, as the analysis here has shown, it turns out that different registers rely on different sets of lexical bundles, and that the bundles have important discourse functions that fit the context and purposes of the registers in which they are common. Their use appears far from accidental; rather, the bundles serve as building blocks of typical discourse within the register.

The fact that most of the lexical bundles are not structurally complete has likely contributed to their being overlooked in previous research, since traditionally linguists have focused on grammatical phrases and clauses, rather than lexical units that cut across grammatical structures. Furthermore, most of the bundles are quite transparent in meaning. As such, they have also been overlooked by researchers who consider idiomaticity a requirement for language that is non-compositional, although there is no reason that semantically transparent sequences could not also be processed as whole chunks (see further Erman/ Warren 2000, 54).

Although the majority of words do not occur within recurrent sequences in either conversation or academic prose, the frequency and functions of lexical bundles demonstrate that speakers and writers use them regularly in building discourse. While much further study is needed—particularly from a psycholinguistic perspective and in more registers—lexical bundles already deserve attention in thorough lexicographic descriptions of English.

Brought to you by | Portland State University (Portland State University)
Authenticated | 172.16.1.226
Download Date | 3/27/12 1:02 AM

7 Bibliography

ALTENBERG 1993 = BENGT ALTENBERG: Recurrent word combinations in spoken English. In: J. D'ARCY: Proceedings of the Fifth Nordic Association for English Studies Conference. University of Island, Reykjavik 1993.

ALTENBERG 1998 = BENGT ALTENBERG: On the phraseology of spoken English: The evidence of recurrent word-combinations. In: A. P. COWIE: Phraseology. Oxford University Press, Oxford 1998, 101–122.

BIBER/CONRAD 1999 = DOUGLAS BIBER/SUSAN CONRAD: Lexical bundles in conversation and academic prose. In: HILDE HASSELARD, SIGNE OKSEFJELL: Out of Corpora: Studies in Honor of STIG JOHANSSON. Rodopi, Amsterdam 1999, 181–189.

BIBER/CONRAD/CORTES 2003 = DOUGLAS BIBER/SUSAN CONRAD/VIVIANA CORTES: Lexical bundles in speech and writing: An initial taxonomy. In: ANDREW WILSON/PAUL RAYSON/TONY MCENERY: Corpus Linguistics by the Lune. Peter Lang, Frankfurt/Main 2003, 71–92.

BIBER/CONRAD/CORTES 2004 = DOUGLAS BIBER/SUSAN CONRAD/VIVIANA CORTES: If you look at … Lexical bundles in university lectures and textbooks. Applied Linguistics 25, 2004, 371–405.

BIBER/JOHANSSON/LEECH/CONRAD/FINEGAN 1999 = DOUGLAS BIBER/STIG JOHANSSON/GEOFFREY LEECH/SUSAN CONRAD/EDWARD FINEGAN: The Longman Grammar of Spoken and Written English. Longman, Harlow 1999.

BUTLER 1997 = CHRISTOPHER S. BUTLER: Repeated word combinations in spoken and written text: Some implications for functional grammar. In: CHR. S. BUTLER, J. H. CONNOLLY, R. A. GATWARD, R. M. VISMANS: A Fund of Ideas: Recent Developments in Functional Grammar. University of Amsterdam, Amsterdam 1997, 60–77.

BUTLER 1998 = CHRISTOPHER BUTLER: Collocational frameworks in Spanish. International Journal of Corpus Linguistics 3, 1998, 1–32.

ERMAN/WARREN 2000 = BRITT ERMAN/BEATRICE WARREN: The idiom principle and the open choice principle. Text 20, 2000, 29–62.

FIRTH 1957 = J. R. FIRTH: Papers in Linguistics, 1934–1951. Oxford University Press, London 1957.

HYMES 1968 = DELL HYMES: The ethnography of speaking. In: J. A. FISHMAN: Readings in the Sociology of Language. Mouton, The Hague 1968, 99–138.

KENNEDY 1993 = GRAEME KENNEDY: Amplifier collocations in the British National Corpus: Implications for English language teaching. TESOL Quarterly 37, 1993, 467–487.

MCCARTHY/CARTER 1997 = MICHAEL MCCARTHY/RONALD CARTER: Written and spoken vocabulary. In: NORBERT SCHMITT, MICHAEL MCCARTHY: Vocabulary: Description, Acquisition and Pedagogy. Cambridge University Press, Cambridge 1997, 20–39.

MOON 1997 = ROSAMUND MOON: Vocabulary connections: Multi-word items in English. In: NORBERT SCHMITT, MICHAEL MCCARTHY: Vocabulary: Description, Acquisition and Pedagogy. Cambridge University Press, Cambridge 1997, 40–63.

MOON 1998a = ROSAMUND MOON: Fixed Expressions and Idioms in English: A Corpus-Based Approach. Clarendon, Oxford 1998.

MOON 1998b = ROSAMUND MOON: Frequencies and forms of phrasal lexemes in English. In: A. COWIE: Phraseology. Oxford University Press, Oxford 1998, 79–100.

NATTINGER/DECARRICO 1992 = JAMES NATTINGER/JEANETTE DECARRICO: Lexical Phrases and Language Teaching. Oxford University Press, Oxford 1992.

PAWLEY/SYDER 1983 = ANDREW PAWLEY/FRANCES SYDER: Two puzzles for linguistic theory: Native-like selection and native-like fluency. In: JACK RICHARDS, RICHARD SCHMIDT: Language and Communication. Longman, London 1983, 191–226.

RENOUF/SINCLAIR 1991 = ANTOINETTE RENOUF/JOHN SINCLAIR: Collocational frameworks in English. In: KARIN AIJMER, BENGT ALTENBERG: English Corpus Linguistics: Studies in Honour of JAN SVARTVIK. Longman, London 1991, 128–143.

SINCLAIR 1991 = J. SINCLAIR: Corpus, Concordance, Collocation. Oxford University Press, Oxford 1991.

Brought to you by | Portland State University (Portland State University)
Authenticated | 172.16.1.226
Download Date | 3/27/12 1:02 AM

WRAY 2002 = ALISON WRAY: Formulaic language and the lexicon. Cambridge University Press, Cambridge 2002.

WRAY/PERKINS 2000 = ALISON WRAY/MICHAEL PERKINS: The functions of formulaic language: an integrated model. Language and Communication 20, 2000, 1–28.

Brought to you by | Portland State University (Portland State University)
Authenticated | 172.16.1.226
Download Date | 3/27/12 1:02 AM