A corpus-driven approach to formulaic language in English*

Multi-word patterns in speech and writing

Douglas Biber
Northern Arizona University

The present study utilizes a corpus-driven approach to identify the most common multi-word patterns in conversation and academic writing, and to investigate the differing pattern types in the two registers. The paper first surveys the methodological characteristics of corpus-driven research and then contrasts the linguistic characteristics of two types of multi-word sequences: 'multi-word lexical collocations' (combinations of content words) versus 'multi-word formulaic sequences' (incorporating both function words and content words).

Building on this background, the primary focus of the paper is an empirical investigation of the 'patterns' represented by multi-word formulaic sequences. It turns out that the multi-word patterns typical of speech are fundamentally different from those typical of academic writing: patterns in conversation tend to be fixed sequences (including both function words and content words). In contrast, most patterns in academic writing are formulaic frames consisting of invariable function words with an intervening variable slot that is filled by content words.

Keywords: corpus-driven research, lexical bundles, lexical patterns, multi-word formulaic sequences, conversation versus academic writing

1. Introduction

Over the past 30 years, there has been increasing interest in the linguistic study of phraseology, identifying different kinds of formulaic multi-word sequences and describing how these sequences are used in natural discourse (see the reviews in Weinert 1995; Ellis 1996; Howarth 1996, 1998; Wray & Perkins 2000; and Wray 2002). Several different approaches have been developed, using different criteria for the identification of multi-word sequences. For example, some studies describe multi-word sequences that are idiomatic (e.g. expressions like *in a nutshell*), while

International Journal of Corpus Linguistics 14:3 (2009), **275–311**. DOI 10.1075/ijcl.14.3.08bib
ISSN 1384–6655 / E-ISSN 1569–9811

other studies focus on sequences that are non-idiomatic but perceptually salient (e.g. *you're never going to believe this*).

More recently, there has been considerable attention given to the ways in which formulaic language can be studied in large text corpora. Two general approaches can be distinguished: 'corpus-based' and 'corpus-driven' (see Tognini-Bonelli 2001: 84–87). 'Corpus-based' research assumes the validity of linguistic structures derived from linguistic theory; the primary research goal is to analyze the systematic patterns of use for those pre-defined linguistic features. Thus, in corpus-based studies of formulaic language, the researcher pre-selects formulaic expressions, and then analyzes the corpus to discover how those expressions are used (e.g. Moon 1998).

In contrast, 'corpus-driven' research is more inductive, so that the linguistic constructs themselves emerge from analysis of a corpus. The corpus-driven approach differs from the standard practice of linguistics in that it makes minimal a priori assumptions regarding the linguistic constructs that should be employed for the analysis. In its most basic form, corpus-driven analysis assumes only the existence of words; co-occurrence patterns among words, discovered from the corpus analysis, are the basis for subsequent linguistic descriptions.

Studies of lexical 'collocations' (see, e.g., Sinclair 1991a, Partington 1998, McEnery et al. 2006) are usually considered to be corpus-driven. Collocational studies often contrast the lexical associations of supposedly synonymous words, such as the differing collocations of *big*, *large*, and *great* (discussed in Biber et al. 1998: 43–53). Studies of this type are corpus-driven in that the lexical collocations of a target word are discovered through corpus analysis. However, there is also a preliminary step in the analysis where the linguist chooses "interesting" target words.

These two steps in the analysis become more important for the study of multi-word formulaic sequences, and as a result, the analysis often ends up combining corpus-based and corpus-driven methodologies. For example, Renouf & Sinclair (1991) undertook a study of multi-word 'collocational frameworks': pairings of function words with a variable lexical slot (e.g., *a + ? + of*, *be + ? + to, many + ? + of*). The Renouf & Sinclair (1991) study is corpus-driven because the sets of lexical fillers were discovered through corpus analysis. At the same time, this study is corpus-based because the collocational frameworks were pre-selected by Renouf and Sinclair.

Another corpus-driven approach to the study of multi-word formulaic sequences makes fewer theoretical assumptions, beginning with simple word forms and using frequency distributions to identify recurrent word sequences (e.g. Salem 1987; Altenberg & Eeg-Olofsson 1990; Altenberg 1998; Butler 1998). Previous studies of 'lexical bundles' in spoken and written registers have adopted this approach (see, e.g., Biber et al. 1999: Chapter 13; Biber et al. 2004).

The present study integrates the above two approaches to identify the most common multi-word patterns (including discontinuous sequences) in conversation and academic writing, and to investigate the differing ways in which those patterns are variable in the two registers. The paper begins with a brief review of previous corpus-driven research on formulaic language, noting that actual practice has sometimes relaxed the methods advocated in theoretical discussions. The paper then compares two specific quantitative methods used for the corpus-driven identification of formulaic sequences: probabilistic statistics (e.g. Mutual Information Score) versus simple frequency. When applied to multi-word sequences (i.e. three or more words), these two methodologies provide very different perspectives on the kinds of formulaic sequences found in a corpus. I discuss statistical reasons for these differences and argue for the existence of two underlying linguistic constructs: 'multi-word lexical collocations' versus 'multi-word formulaic sequences'. The former are often multi-word technical terms, consisting of a sequence of lexical words. In contrast, the latter are usually high-frequency sequences that include both function words and content words. These sequences can be classified according to their general 'patterns', with fixed and variable slots occurring in a particular order.

The remainder of the paper presents an empirical investigation of the multi-word patterns in speech and writing. The study adopts a radical corpus-driven approach to investigate the ways in which high-frequency sequences of words pattern in terms of fixed and variable slots. First, the full set of multi-word sequences that occur most frequently in the corpus is identified. Each of those multi-word sequences is then investigated to describe its variable properties: the extent to which each slot in the sequence is fixed or variable. This analysis empirically distinguishes among several types of multi-word 'patterns', including multi-word sequences with an internal variable slot (similar to 'collocational frameworks'), sequences with multiple variable slots, and relatively fixed lexical sequences.

A major innovation of the present study is that it incorporates a register perspective, in contrast to most previous studies of collocation and formulaic language, which have disregarded register differences. The findings show that this perspective is essential for studies of formulaic language, because both the specific patterns and the more general pattern types commonly used in spoken discourse are fundamentally different from the common pattern types in written discourse.

2. Methodological characteristics of previous corpus-driven research

2.1 The role of pre-defined grammatical categories in corpus-driven research

A corpus-based approach is the norm for studies of grammatical variation (see, e.g., Biber et al. 1999; Lindquist & Mair 2004; Rohdenburg & Mondorf 2003); these studies usually assume the grammatical categories defined in previous linguistic research (e.g. passive voice, progressive aspect, nominalizations, relative clauses), and then use corpus research to describe the patterns of variation and use associated with those grammatical features.

In contrast, many studies of collocation and formulaic language have been corpus-driven, exploiting the potential of a corpus to identify linguistic categories and units that have not been previously recognized. That is, in a corpus-driven analysis, the "descriptions aim to be comprehensive with respect to corpus evidence" (Tognini-Bonelli 2001: 84), so that even the 'linguistic categories' are derived "systematically from the recurrent patterns and the frequency distributions that emerge from language in context" (Tognini-Bonelli 2001: 87).

In its most extreme form, the corpus-driven approach assumes only the existence of word forms; grammatical classes and syntactic structures have no a priori status in the analysis. In fact, even inflected variants of the same lemma are treated separately, with the underlying claim that each word form has its own grammar and its own meanings. (See, for example, the discussion of *eye* versus *eyes* in Sinclair 1991b and Stubbs 1993: 16.)

In actual practice, a fairly wide range of methodologies have been used under the umbrella of corpus-driven research. These methodologies can all be distinguished from pure corpus-based research by the nature of their central research goal: to uncover new linguistic constructs through inductive analysis of corpora. However, corpus-driven studies often incorporate some corpus-based methods, and thus they might be best considered as hybrids.

For example, one major application of corpus-driven research is the 'pattern grammar' framework, referred to as *A Corpus-Driven Approach to the Lexical Grammar of English* (Hunston & Francis 2000). Two major reference books have emerged from this framework (Francis et al. 1996; 1998). These studies are corpus-driven in that they focus primarily on the construct of the grammatical pattern: "a phraseology frequently associated with (a sense of) a word…" (Hunston & Francis 2000: 3). These books show that there are systematic regularities in the associations between grammatical frames, sets of words, and particular meanings on a much larger scale than it could have been possible to anticipate before the introduction of large-scale corpus analysis. Grammatical patterns are not necessarily complete structures (phrases or clauses) recognized by linguistic theory. Thus, following the

central defining characteristic of corpus-driven research given above, the pattern grammar studies attempt to uncover new linguistic constructs — the patterns — through inductive analysis of corpora.

The pattern grammar studies are instructive here because they are often cited as the best developed example of corpus-driven research, but in practice they employ both corpus-driven and corpus-based methodologies. The studies are corpus-driven because the lexical associations of each pattern are discovered through corpus analysis. However, the studies are corpus-based because the analyses are in part determined by pre-defined linguistic categories (including basic grammatical categories like 'noun' and 'verb', phrase types, and even syntactic structures).

For example, the initial analyses for the pattern grammar reference books involve a classification of patterns according to three major part of speech categories (nouns, adjectives, and verbs). Thus, patterns are not simply discovered through corpus analysis of word forms. Rather, they are combinations of pre-defined grammatical categories co-occurring with particular words (e.g. **ADJ *in* N**). (See further the discussion in Mahlberg 2005: 75–77.)

For some patterns, the methods are even more fundamentally corpus-based in that they require a priori syntactic analysis. For example, the pattern **N that** is defined so that it includes only 'appositive clauses' (Hunston & Francis 2000: 98–99; Francis et al. 1998: 108–113). Thus, this pattern includes only nouns that are followed by an appositive *that*-clause (e.g. *fact, claim, stipulation, expectation, disgust, problem*, etc.). In contrast, nouns followed by the relative pronoun *that* do not constitute any pattern, even though corpus analysis shows that some of these combinations are quite frequent (e.g. *extent*, *way*, *thing*, *questions*, *evidence*, *factors* + *that*).

Similarly, the pattern **ADJ in N** is defined to include only prepositional phrases that complement the adjective, such as *involved in*, *deficient in*, *rich in*, *proficient in*. In contrast, other frequent combinations of **ADJ in N** are excluded because the prepositional phrase has an adverbial function (see Francis et al. 1998: 444–451; Hunston & Francis 2000: 75–76). As a result, combinations such as *firm in*, *resolute in*, and *steadfast in* are excluded from the analysis and do not belong to any pattern.

The theoretical position adopted here is that corpus-driven research is not in any way superior to corpus-based research. But the two require radically different methods, and thus they offer the possibility of uncovering radically different perspectives on language structure and use. However, to realize that potential, the approaches must be clearly distinguished in practice and applied systematically.

In sum, studies that claim to be corpus-driven do not necessarily adopt the strict methods advocated by Sinclair (1991a, b), assuming only the existence of word forms with no reference to grammatical categories (or even lemmas). The

hybrid approach adopted by the pattern grammar studies — incorporating both corpus-based and corpus-driven elements — has proven to be highly informative. At the same time, it should also be useful to explore the potential of a radical corpus-driven approach to multi-word formulaic sequences.

2.2 The role of frequency in corpus-driven research

The role of frequency and quantitative analysis in corpus-driven research is even more controversial. Nearly every theoretical description of the corpus-driven approach includes mention of frequency, as in:

> the "linguistic categories" are derived "systematically from the recurrent patterns and the frequency distributions that emerge from language in context" (Tognini-Bonelli 2001: 87)

> in a grammar pattern, "a combination of words occurs relatively frequently" (Hunston & Francis 2000: 37)

In actual practice, though, frequency is not important in many studies claiming to be corpus-driven. For example, most pattern grammar studies report no frequency findings, and there is no evidence that frequency was actually used in the analyses. In fact, frequent word combinations are explicitly omitted from the pattern analysis if they represent different syntactic constructions (as described above).

Surprisingly, some corpus-driven linguists have overtly argued against the importance of frequency. For example, Sinclair notes that

> some numbers are more important than others. Certainly the distinction between 0 and 1 is fundamental, being the occurrence or non-occurrence of a phenomenon. The distinction between 1 and more than one is also of great importance … [because even two unconnected tokens constitute] the recurrence of a linguistic event…, [which] permits the reasonable assumption that the event can be systematically related to a unit of meaning. In the study of meaning it is not usually necessary to go much beyond the recognition of recurrence [i.e. two independent tokens]….
>
> (Sinclair 2001: 343–4)

Similarly, Tognini-Bonelli notes that

> It is therefore appropriate to set up as the minimum sufficient condition for a pattern of occurrence to merit a place in the description of the language, that it occurs at least twice, and the occurrences appear to be independent of each other….
>
> (Tognini-Bonelli 2001: 89)

Thus, there is a tension here between the underlying definition of the corpus-driven approach, which derives linguistic categories from 'recurrent patterns' and 'frequency distributions' (Tognini-Bonelli 2001: 87), and the actual practice of scholars working on pattern grammar and the lexis-grammar-meaning interconnection, which has focused much more on form-meaning associations with relatively little accountability to quantitative distributional patterns in a corpus.

2.3 Characteristics of a strict corpus-driven approach

Synthesizing previous theoretical discussions, a radical corpus-driven approach to formulaic language would have three general characteristics:

1. it would be based on analysis of the actual word forms that occur in the corpus (not lemmas)
2. it would be based on analysis of sequences of word forms, with no consideration given to the grammatical/syntactic status of those words
3. it would focus on frequent, recurrent combinations of word forms

Collocational studies have generally adopted these characteristics (e.g. Sinclair 1991a, b; Partington 1998; Stubbs 1993, 1995; Mahlberg 2005), including a focus on the word combinations that are recurrent rather than word combinations that occur only once or twice in a corpus. However, most studies of longer multi-word sequences or lexico-grammatical combinations have not strictly adhered to these guidelines, even when they claim to be corpus-driven.

It is important to emphasize the main point here. The discussion above does not challenge the value of previous hybrid studies that have claimed to be corpus-driven. Rather, the point here is simply that there are clear methodological distinctions between corpus-based and corpus-driven approaches, and that the distinction should be applied in practice. Both approaches are valuable (as is a hybrid blending of the two approaches). But at the same time, it is useful to be clear about the actual methods that have been adopted, and to explore the analytical potential of all approaches.

The question that the present paper explores is whether a radical corpus-driven approach can be applied to the study of formulaic language. In particular, the approach taken in the following sections incorporates all three defining characteristics of corpus-driven research: analyzing word forms rather than lemmas; considering only sequences of words, regardless of the pre-defined grammatical categories; and using frequency information derived from the corpus as the primary evidence to be considered in the analysis.

2.3.1 *An example: Previous research on lexical bundles*

Although they have generally not adopted the term 'corpus-driven research', several researchers interested in the study of formulaic language have adopted a radical corpus-driven approach, beginning with simple word forms and giving priority to frequency to identify recurrent word sequences (e.g. Salem 1987; Altenberg & Eeg-Olofsson 1990; Altenberg 1998; Butler 1998). The psycholinguistic status of recurrent word sequences has also been investigated by Schmitt et al. (2004) and Ellis et al. (2008).

More recently, such research has been carried out using the construct of 'lexical bundle': the most frequently recurring sequences of words in a register (e.g. *I don't know if, I just wanted to*). The term 'lexical bundle' was first used in the Longman Grammar of Spoken and Written English (Biber et al. 1999: Chapter 13), and has since then been applied in several subsequent studies, including Biber et al. (2004), Biber & Barbieri (2007), Cortes (2004), Nesi & Basturkmen (2006), and Partington & Morley (2004).

Lexical bundles are identified using a frequency-driven approach. In the initial study of English lexical bundles (Biber et al. 1999: Chapter 13), a relatively low frequency cut-off was used: 10 times per million words. However, a sub-set of these bundles, occurring more than 40 times per million words, was used for detailed analyses of structural characteristics and discourse functions. Many of these bundles are actually much more common, occurring more than 200 times per million words.

Lexical bundles of any length can be analyzed. For example, the initial description of English bundles was based on 3-word, 4-word, and 5-word sequences, but only 4-word sequences were considered in the more detailed analyses. A further defining characteristic is that a multi-word sequence must be used in multiple texts to be counted as a lexical bundle (at least five different texts), to guard against idiosyncratic uses by individual speakers or authors. Most bundles are distributed widely across the texts in a corpus. For example, even the least common lexical bundles in conversation or classroom teaching are usually used in at least 20 different texts.

The initial analysis of lexical bundles in English (Biber et al. 1999: Chapter 13) compared the patterns of use in conversation and academic prose, based on analysis of c. 5-million-word sub-corpora from each register. Figure 1 shows the overall distribution of all lexical bundles occurring more than 10 times per million words (distributed across at least five different texts). Not surprisingly, there are almost ten times as many 3-word bundles as 4-word bundles. It is perhaps more surprising that there are many more lexical bundles in conversation than in academic writing, and this pattern is even stronger for the longer bundles.

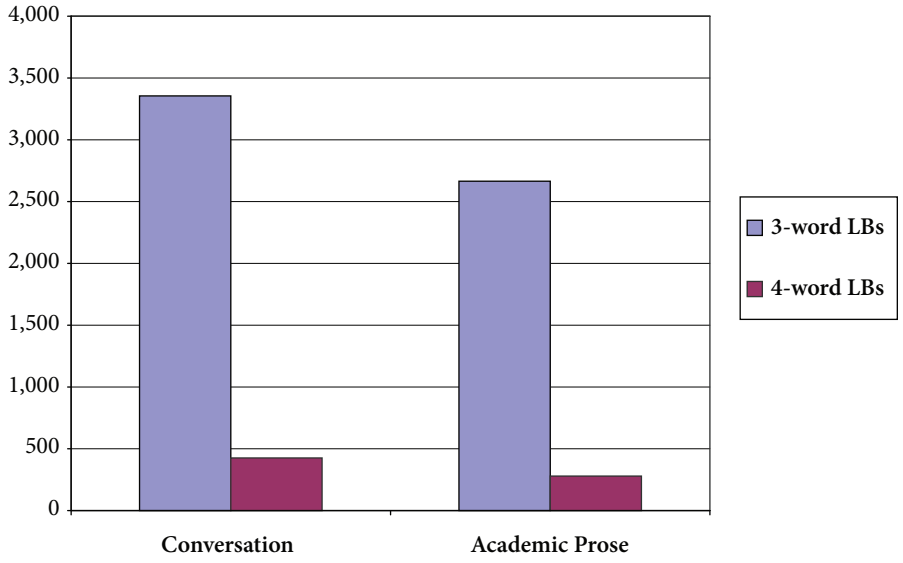

Figure 1. Number of different lexical bundles in conversation versus academic prose (occurring more than 10 times per million words)

Although the set of lexical bundles is identified based only on distributional criteria (rate of occurrence and distribution across texts), it turns out that bundles tend to have several characteristics that distinguish them from other kinds of formulaic expressions. First, lexical bundles are by definition extremely common (in contrast to most idioms and many 'grammar patterns', which tend to be rare). Second, most lexical bundles are not idiomatic in meaning and not perceptually salient. For example, the meanings of bundles like *do you want to* or *I don't know what* are transparent from the individual words. And finally, lexical bundles usually do not represent a complete structural unit. For example, Biber et al. (1999: 993–1000) found that only 15% of the lexical bundles in conversation can be regarded as complete grammatical phrases or clauses, while less than 5% of the lexical bundles in academic prose represent complete structural units. 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 beginning elements of a second grammatical structure. Most of the bundles in speech bridge two clauses (e.g. *I want to know, well that's what I*), while bundles in writing usually bridge two phrases (e.g. *in the case of, the base of the*).

Although they are not complete structural units, bundles do usually have strong grammatical correlates. For example, bundles like *you want me to* are constructed from clause components, while bundles like *in the case of* are constructed from phrase components. Many clausal bundles simply incorporate verb phrase frag-

ments, such as *it's going to be* and *what do you think*. Other clausal bundles are composed of dependent clause fragments, such as *when we get to* and *that I want to*. In contrast, phrasal bundles either consist of noun phrase components, usually ending with the start of a postmodifier (e.g. *the end of the*, *those of you who*), or prepositional phrase components with embedded modifiers (e.g. *of the things that*).

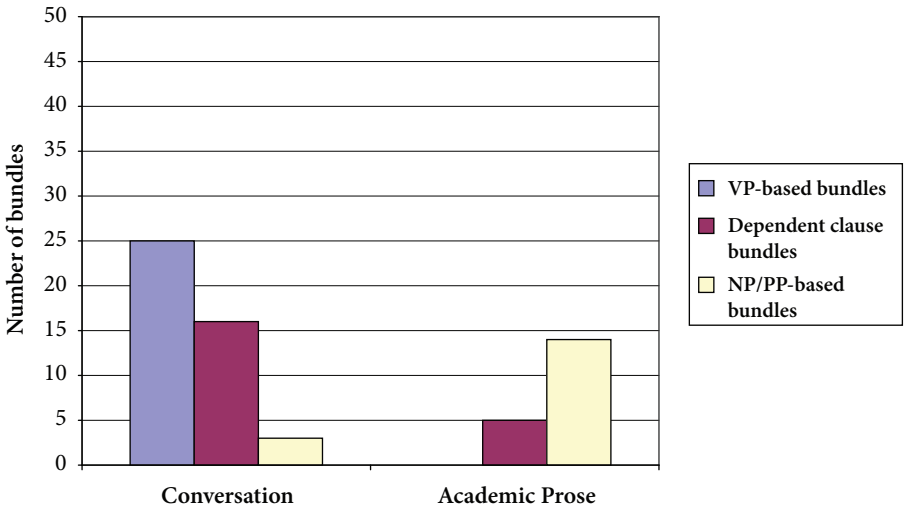

Figure 2. Distribution of common lexical bundles across structural types (4-word bundles occurring more than 40 times per million words)

Figure 2 plots the distribution of these lexical bundle types across registers, showing that the structural correlates of lexical bundles in conversation are strikingly different from those in academic prose. In conversation, c. 90% of all common lexical bundles are declarative or interrogative clause segments (i.e. 41 out of a total of 44 bundles that occur more than 40 times per million words). In fact, c. 50% of these lexical bundles begin with a personal pronoun + verb phrase (such as *I don't know why, I thought that was*). In contrast, most lexical bundles in academic prose are phrasal rather than clausal. There were no common VP-based bundles in this corpus of academic prose, and only five common bundles that incorporate dependent clauses (e.g. *the way in which, the fact that the, it is possible to*). In contrast, 14 of the 19 common bundles in academic prose (74%) 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*).

It further turns out that lexical bundles in English have systematic discourse functions. That is, although they are neither idiomatic nor structurally complete, lexical bundles are important building blocks in discourse. Lexical bundles provide a kind of pragmatic 'head' for larger phrases and clauses, where they func-

tion as discourse frames for the expression of new information. That is, the lexical bundle expresses stance or textual meanings, while the remainder of the phrase/clause expresses new propositional information that has been framed by the lexical bundle. In this way, lexical bundles provide interpretive frames for the developing discourse. For example,

(1) ***I want you to*** *write a very brief summary of his lecture.*

(2) *Hermeneutic efforts are provoked by* ***the fact that the*** *interweaving of system integration and social integration […] keeps societal processes transparent…*

Biber et al. (2004) identified three primary discourse functions for lexical bundles in English: 1) stance expressions, 2) discourse organizers, and 3) referential expressions. Stance bundles express epistemic evaluations or attitudinal / modality meanings:

(3) ***I don't know what*** *the voltage is here.*

(4) ***I don't want to*** *deliver bad news to her.*

Discourse organizers reflect relationships between prior and coming discourse: introducing topics, topic elaboration/clarification, confirmation checks, etc.:

(5) ***What I want to do is*** *quickly run through the exercise…*

(6) *Yes,* ***you know there was*** *more of a playful thing with it,* ***you know what I mean****?*

Finally, referential bundles identify an entity or single out some particular attribute of an entity as especially important:

(7) *Students must define and constantly refine* ***the nature of the*** *problem.*

(8) *She's in that office down there,* ***at the end of the*** *hall.*

Figure 3 shows that the typical functions of lexical bundles are strikingly different in conversation versus academic writing: most bundles are used for stance functions in conversation, with a number also being used for discourse organizing functions. In contrast, most bundles are used for referential functions in academic prose. These findings indicate that formulaic expressions develop to serve the most important communicative needs of a register.

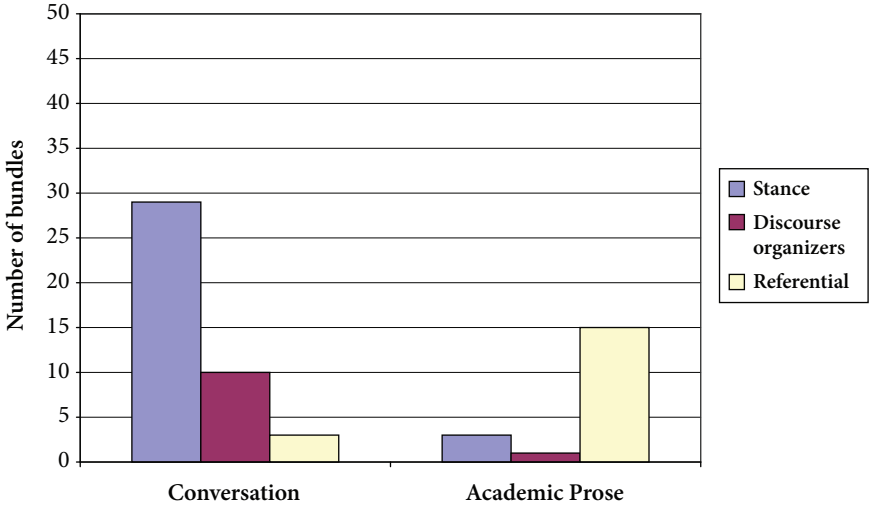

Figure 3. Distribution of common lexical bundles across functional types (4-word bundles occurring more than 40 times per million words)

It is worth noting that only the initial identification of lexical bundles and the analyses of their distributions employ corpus-driven methods. In contrast, the subsequent interpretations — both structural and functional — employ constructs from linguistic theory.

3. Multi-word collocations versus multi-word formulaic sequences

One methodological issue that arises in corpus-driven studies of formulaic language concerns the best way to determine whether a word combination is recurrent. That is, researchers on collocation have observed that absolute frequency often fails to capture the word associations that are most important for lexical research (see, e.g., Oakes 1998; McEnery et al. 2006: 208–220).

For example, a search on the word *dog* in a 4-million word corpus of fiction shows that it most frequently co-occurs with words like *the*, *a*, *his*, *that*. However, these co-occurrences are not very interesting from a collocational perspective, since they represent grammatical patterns. Restricting the focus to preceding adjectives, we still find that simple frequency does not necessarily capture the strongest collocational associations. For example, the adjectives *old* and *big* are the most frequent premodifiers of *dog* in this corpus. However, those adjectives frequently co-occur with many different nouns, and so are not especially associated with the noun *dog*.

An alternative approach is to use a statistical measure of association, like the Mutual Information (MI) score. These scores compare the frequency of a word combination to the overall frequencies of each of the individual words; as the individual word frequencies become higher, it becomes more likely that the word combination would occur just by random chance, and therefore the combination has less importance as a collocation. MI scores reflect this collocational strength, as in:

	frequency of word combination	**MI score**	**frequency of the first word**
his dog	12	1.6	> 37,000
old dog	8	4.1	4,233
stray dog	2	9.1	34

Thus, although the combination *old dog* is relatively frequent, only about 0.2% of the tokens of *old* occur together with *dog* (8/4,233). In contrast, the combination *stray dog* is less frequent, but *dog* is one of the few nouns that *stray* co-occurs with: 6% of the tokens of *stray* co-occur with *dog* (2/34). (This relationship is much stronger if we exclude occurrences of *stray* as a verb.) For these reasons, statistical measures like MI score are commonly used to evaluate the collocational strength of association between words, in contrast to simple frequency of occurrence for a word combination.

In recent years, this same approach has been applied to multi-word sequences. Thus, software programs like Collocate (Barlow 2004) automatically compute an MI score for extended multi-word sequences, and this methodology has been applied in empirical studies of formulaic sequences like Ellis et al. (2008).

However, there are two concerns about this application of the MI statistic that deserve discussion. First, because the statistic was developed to measure the collocational strength of word pairs, it does not take into account the order of words. Rather, the statistic is based on a simple comparison of the word-pair frequency with the frequency that would be predicted by chance. The predicted frequency is based on the actual frequency of each word in isolation, compared to the total size of the corpus.

The main point here is that the predicted frequency is identical regardless of the word order. For example, we would predict that *old dog* should occur .47 times per million words just by random chance:

$$[\ 4{,}233\ (\text{freq. of } \textit{old}) \times 446\ (\text{freq. of } \textit{dog})\ ] \ /\ 4{,}000{,}000\ (\text{total corpus}) = .47$$

However, we would also predict that *dog old* should occur with exactly the same frequency:

[446 (freq. of *dog*) × 4,233 (freq. of *old*)] / 4,000,000 (total corpus) = .47

The MI score is based on this predicted frequency, simply reflecting the likelihood that the two words occur together, regardless of order. This characteristic poses no problem in practice for the analysis of two-word combinations. However, as shown below, this characteristic becomes problematic when applied to multi-word sequences, where part of the formulaic nature of the sequence is the fixed word order itself.

An even more basic issue concerns the linguistic goal of the analysis. As noted above, the MI statistic disfavors combinations that incorporate high-frequency words, to deliberately exclude function words from consideration. The underlying assumption is that collocations are combinations of content words, while a lexico-grammatical combination of function word plus content word is a different phenomenon (e.g. *stray dog* versus *the dog*). When this assumption is extended to the operational definition of formulaic multi-word sequences, it results in the conclusion that sequences of content words are formulaic, while sequences that include function words are simply grammatical by-products. But in fact there are different kinds of multi-word formulaic expressions in English, and different quantitative methods are required to capture them.

To illustrate, the program Collocate was used to extract the 4-word sequences with the highest MI scores in a corpus of academic prose (the 5.3 million-word sub-corpus of academic prose included in the LSWE Corpus; see Biber et al. 1999: 24ff). For the first pass, the search was restricted to word sequences that occurred at least 10 times in the corpus. In this case, all of the sequences with the highest MI scores were technical terms, consisting entirely of lexical words; for example:

4-word sequence	**Frequency**	**MI score**
trinitro benzene sulphonic acid	10	46.6
giardia heat shock antigen	10	39.6
lamina propria mononuclear cells	23	38.8
static torque rotor position	28	35.5
first lateral arm plate	67	33.5

In a second pass, Collocate was used to compute the MI scores for the most frequent 4-word sequences in this corpus of academic prose (identified in a previous study of lexical bundles; Biber et al. 1999: Ch. 13). Table 1 lists the lexical bundles that have the largest MI scores.

Table 1. Frequencies and MI scores for selected lexical bundles in academic prose

Lexical bundle	Lexical bundle freq.	Freq. of word 1	Freq. of word 2	Freq. of word 3	Freq. of word 4	MI score
it should be noted	112	36,773	6,446	47,919	827	21.0
as we have seen	138	41,581	15,434	21,490	2,678	19.3
on the other hand	550	28,188	366,393	10,929	1,352	19.3
at the same time	377	21,766	366,393	4,645	6,695	18.1
on the one hand	157	28,188	366,393	14,297	1,352	17.1
it is clear that	128	36,773	85,154	1,445	55,848	16.3
can be used to	176	16,549	47,919	6,330	132,533	15.5
are likely to be	118	40,028	1,952	132,533	47,919	15.3
it is necessary to	211	36,773	85,154	2,320	132,533	15.3
it is possible to	335	36,773	85,154	3,764	132,533	15.2
…						
the extent to which	219	366,393	1,195	132,533	27,114	14.5
to be able to	174	132,533	47,919	1,725	132,533	14.3
…						
in the case of	506	131,769	366,393	3,553	227,311	11.2
the end of the	367	366,393	1,699	227,311	366,393	10.2
the nature of the	289	366,393	1,683	227,311	366,393	9.9

By definition, these bundles have much higher frequencies: occurring over 100 times in this 5-million word corpus (i.e. at least 20 times per million words). But these frequent bundles have much lower MI scores than the 4-word sequences composed of lexical words. All of the high-frequency 4-word sequences incorporate function words; in many cases, three of the four words in the sequence are function words.

A comparison of the two sets of word-sequences illustrates the different kinds of associations identified by the MI approach versus the frequency approach: multi-word sequences with high MI scores tend to be technical referring expressions (usually extended noun phrases) composed of lexical/content words; these can be regarded as **multi-word collocations**. In contrast, the most frequent word sequences (lexical bundles) usually incorporate both function words and lexical words; these can be regarded as **multi-word formulaic sequences**.

A more detailed consideration of the high-frequency 4-word sequences in Table 1 further illustrates the influence of individual word frequency on MI score, and the mismatch between MI score and formulaic status. For example, the sequence

it should be noted has the highest MI score, because all four individual words have relatively low frequencies. In contrast, sequences like *in the case of* and *the end of the* have comparatively low MI scores: because the individual words (especially *in*, *the*, *of*) have extremely high frequencies (occurring over 100,000 times), there is a higher probability that we would find these four words co-occurring by chance (disregarding word order), resulting in a low MI score.

Here again we see that the MI statistic fails to capture the formulaic status of these function word sequences. It is not sufficient to measure the likelihood that four words will co-occur in any order. Rather, the question is the likelihood that these particular words occur in this particular order: how fixed is this exact sequence of words? For example, 82% of the occurrences of the 3-word sequence *the case of* are preceded by the preposition *in*. In contrast, the sequence *on the case of* is unattested in this corpus, even though *on* is also a high frequency preposition. Similarly, 88% of the occurrences of the 3-word sequence *in the case* are followed by the preposition *of*, while the sequence *in the case for* occurs only one time in the corpus. The 4-word sequence *in the case of* is thus strongly formulaic, despite the fact that three of the words in this sequence are high-frequency function words. Simple frequency is better able to capture patterns of this type than the MI statistic.

These explorations lead to the conclusion that collocational strength is not equivalent to formulaic status. Collocational strength characterizes the relationship among lexical words. In contrast, formulaic status incorporates both lexical and function words, occurring together in relatively fixed relationships. Thus, there are different kinds of multi-word sequences, and different quantitative methods are needed to identify them. Based on such considerations, it is possible to rank multi-word associations along a cline of phenomena:

Multi-word collocations ←——————→	**Multi-word formulaic sequences**
sequences of lexical words	sequences of lexical and function words
multi-word technical terms	discourse frames
high MI scores	lower MI scores
lower frequency	high frequency

These should not be considered as two discrete types of formulaic language, but rather as two poles defining a continuum.

An additional consideration is that multi-word formulaic sequences are not necessarily continuous fixed sequences of words. Rather, as described by Renouf & Sinclair (1991), English also employs discontinuous frameworks, combining fixed slots filled by a single function word with variable slots filled by many dif-

ferent content words (e.g. *a/n + ? + of + the*). The following section takes up this perspective.

4. A corpus-driven analysis of fixed and variable formulaic patterns

Both of the quantitative measures discussed above (MI score and simple frequency) — and association statistics generally — are similar in that they evaluate the strength of association for continuous sequences of words. However, as Renouf & Sinclair (1991) suggested almost two decades ago, formulaic language patterns in English can be discontinuous, with high-frequency function words as fixed elements co-occurring with variable lexical slots (e.g. *a/n + ? + of + the; be + ? + to; too + ? + to; many + ? + of*). Renouf & Sinclair (1991) approached this topic using a corpus-based approach, pre-selecting a small number of discontinuous sequences that they had noticed, and then studying those sequences in a corpus to identify the lexical words that filled the variable slots. The question taken up in the remainder of this paper is whether this phenomenon can be studied from a corpus-driven perspective: identifying the patterns that occur most commonly in a corpus; determining the different ways in which those patterns are variable or fixed; and contrasting the overall patterns of use in speech and writing.[1]

Two corpora were analyzed for the study: a 4.5 million word corpus of AmE conversation, and a 5.3 million word corpus of academic prose (research articles and academic books). Both sub-corpora are part of the larger Longman Corpus Network (see Biber et al. 1999: 24–35).

The first step in the analysis was to identify all recurrent multi-word sequences in these two corpora. This step had been carried out previously, as part of the study of lexical bundles (see 2.3.1 above). The goal at that stage was to be as inclusive as possible, providing the basis for the next step: investigation of the different multi-word patterns in speech and writing. Thus, a relatively low frequency cut-off was used for the study: to be included, a sequence had to recur at least 10 times per million words, in at least 5 different texts. To limit the scope of the investigation, only 4-word sequences were considered. (However, the same methods could be employed for the study of both shorter and longer multi-word patterns.)

As shown in previous research (e.g. Biber et al. 1999: 993–5; Biber et al. 2004), conversation employs a larger stock of recurrent multi-word sequences than academic writing. In the corpora analyzed for the present study, 140 multi-word sequences in conversation occurred more than 10 times per million words (see Appendix I), compared to 94 multi-word sequences in academic writing (see Appendix II).

Table 2. Frequencies, MI scores, and the predictive status of each word slot for selected lexical bundles in academic prose

Lexical bundle	Count	MI Score	Slot 1 %	Slot 2 %	Slot 3 %	Slot 4 %	Pattern Type
it should be noted	112	21.0	84	70	94	31	123*
as we have seen	138	19.3	57	96	98	64	1234
on the other hand	550	19.3	97	97	75	78	1234
it is clear that	128	16.3	98	70	9	76	12*4
can be used to	176	15.5	42	97	24	41	*2**
are likely to be	118	15.3	21	26	98	44	**3*
it is necessary to	211	15.3	80	83	10	75	12*4
it is possible to	335	15.2	98	84	16	61	12*4
…							
the extent to which	219	14.5	95	55	99	97	1234
to be able to	174	14.3	30	97	22	90	*2*4
…							
in the case of	506	11.2	82	95	8	88	12*4
the end of the	367	10.2	68	2	95	53	1*34
the nature of the	289	9.9	65	1	97	51	1*34

In the present study, each multi-word sequence was analyzed to identify the patterns of variability. That is, each 'slot' in the sequence was analyzed to determine the extent to which it is variable or fixed. To illustrate, Table 2 lists the same multi-word sequences as Table 1, but shows the extent to which each of the slots is variable: the proportion of the total occurrences of the associated 3-word-combination that is accounted for by a particular word occurring in the 4th slot.

For example, the 3-word sequence * *the other hand* occurred 567 times in the corpus, and 550 of those occurrences — or 97% of the time — were preceded by the word *on*. The 3-word combination *on* * *other hand* also occurred 567 times in the corpus, and in 550 of those — or 97% of the time — the second word was *the*. The last two slots of this sequence are somewhat less fixed. Thus, the 3-word combination *on the* * *hand* occurred 733 times in the corpus, and the third word was *other* in 550 of those occurrences (75% of the time). And finally, the 3-word combination *on the other* * occurred 705 times in the corpus, and in 550 of those occurrences — or 78% of the time — the following word was *hand*.

Table 2 shows that some of these 4-word sequences (e.g. *on the other hand*) are quite formulaic, with all four slots being relatively fixed. In contrast, other sequences are much more variable; for example, in the sequence *are likely to be*, only

21% of the total occurrences of the 3-word sequence * *likely to be* are accounted for by the preceding word *are*. Similarly, only 26% of the occurrences of the 3-word combination *are* * *to be* are accounted for by the word *likely*.

Appendices I and II list the full set of common 4-word sequences in conversation and academic prose, organized by frequency (the rate per one million words of text). These appendices show the variability of each multi-word sequence: the proportion of the associated 3-word-combination that is accounted for by a particular word occurring in each of the slots (similar to Table 2).

The findings presented in Appendices I and II show that there is a continuous space of variation in the extent to which a common 4-word sequence is fixed or variable. Further, these sequences are variable in very different ways. For example:

Variability pattern	Example multi-word sequences
Relatively fixed in all 4 slots	*you want me to, on the other hand*
Relatively fixed in Slots 1, 2, 3 but variable in Slot 4	*I don't know if, it should be noted*
Relatively fixed in Slots 1, 2, 4 but variable in Slot 3	*in the case of, on the basis of, as a result of*

Each 4-word sequence was coded for its 'pattern type', to further analyze the overall preferences for one or another of these variability patterns, as well as potential differences between speech and writing. For this purpose, a simple cut-off of greater or lesser than 50% was used for each slot in a 4-word sequence. That is, if more than 50% of the associated 3-word-combination is accounted for by the particular word occurring in a slot, then that slot is categorized as relatively fixed; otherwise, the slot is categorized as relatively variable. The asterisk is used to represent variable slots. For example, in Table 2, the 4-word sequence *on the other hand* is pattern type 1234: all four slots are relatively fixed. The sequence *in the case of* is pattern type 12*4: the first, second, and fourth slots are relatively fixed, but the third slot is variable.

Figure 4 presents the distribution of multi-word sequences across pattern types, and also provides a comparison of conversation and academic prose. For example, 7% of the common multi-word sequences in conversation are continuous fixed sequences: Pattern Type 1234 (i.e. 10 of the 144 multi-word sequences listed in Appendix I).

As Figure 4 shows, these high-frequency multi-word sequences represent the full spectrum of pattern types. Some of these sequences are relatively fixed in all four slots (e.g. *on the other hand*), while some have no fixed elements at all (e.g., *you can do it*). The majority have a mix of fixed and variable elements but in very

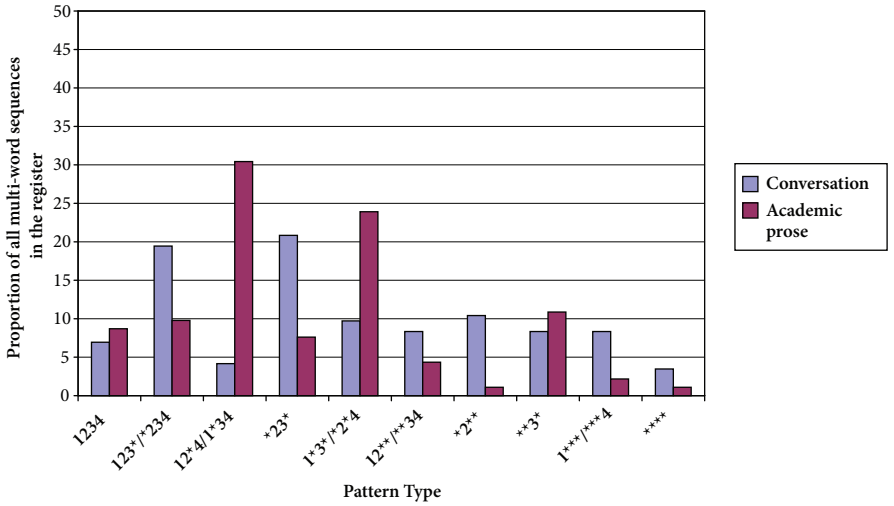

Figure 4. Distribution of common multi-word sequences across pattern types, showing the % of all sequences in conversation vs. academic writing

different configurations. For example, several of these pattern types represent continuous sequences of fixed elements, with variable slots at the beginning or end (e.g. Pattern Type 123*, as in *it should be noted*; Pattern Type 12**, as in *I don't think you*; or Pattern Type **34, as in *is equal to the*). In contrast, several other pattern types consist of fixed elements with an intervening variable slot (e.g. Pattern Type 1*34, as in *the end of the*; or Pattern Type 12*4, as in *in the case of*).

Conversation and academic writing both use the full range of pattern types. However, as Figure 5 shows, there are strong differences in the preferred pattern types of the two registers. In academic writing, over 50% of the multi-word sequences represent formulaic frames with an internal variable slot (i.e. two or three fixed elements with 1–2 intervening variable slots, as in 1*34, 12*4, 1*3*, etc.). In contrast, over 50% of the multi-word sequences in conversation represent continuous sequences of fixed elements, with a preceding or following variable slot (e.g. *234, 12**, etc.). In addition, a high proportion (over 30%) of the high-frequency multi-word sequences in conversation do not represent fixed patterns, consisting of only one fixed element co-occurring with three variable slots (or even four variable elements for a few sequences).

It further turns out that the elements filling fixed and variable slots are fundamentally different in speech and writing. In conversation, most 4-word sequences are composed of one content word in sequence with three function words (e.g. *I don't know if*). However, Figure 6 shows that both content words and function words are equally likely to be fixed in conversation. Stated the other way around, both fixed and variable slots are usually filled by function words in conversation.

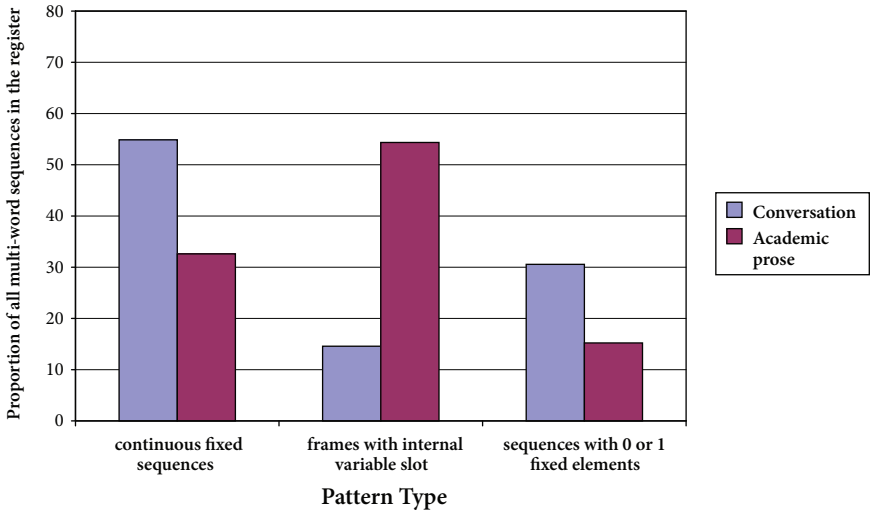

Figure 5. Distribution of common multi-word sequences in conversation vs. academic writing: Continuous sequences (1234, 123*, *234, *23*, 12**, **34) versus frames with internal variable slots (1*34, 12*4, 1*3*, *2*4, 1**4)

In contrast, Figure 6 shows a dramatic difference between fixed slots and variable slots in academic prose: fixed slots are almost always filled by content words, while variable slots are almost always filled by function words.

Finally, these findings can be explored from the opposite perspective, identifying the patterns that occur most frequently. Previous research on 'lexical bundles'

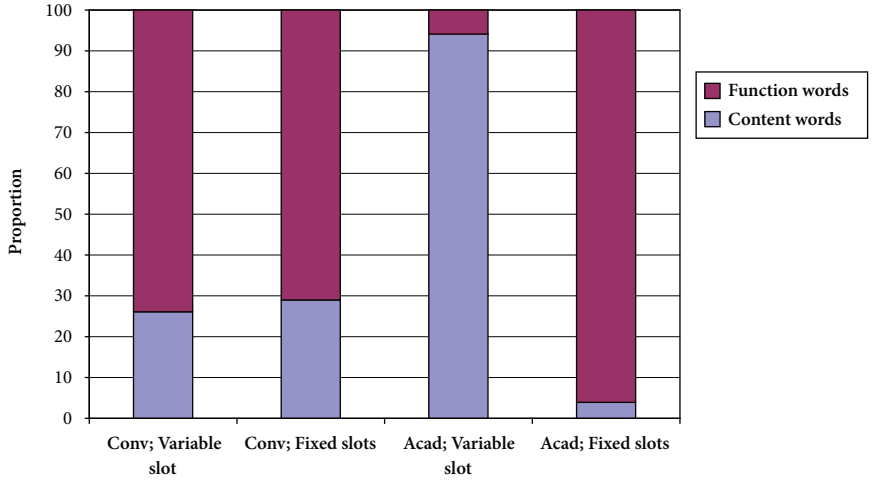

Figure 6. Structural characteristics of the fillers in fixed slots versus variable slots (for multi-word sequences with one variable slot)

has shown that high-frequency multi-word sequences are more prevalent in conversation than in academic writing (see Section 2.3.1 above). The most common individual lexical bundles also occur with higher frequencies in conversation than in writing, as shown in Figure 7. In contrast, specific patterns have the opposite

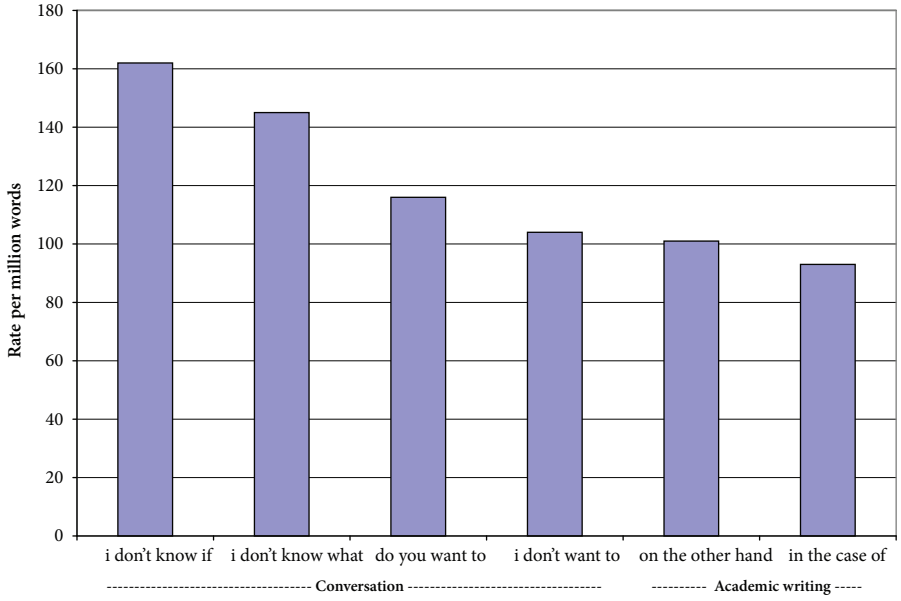

Figure 7. The most frequent 4-word sequences in conversation and academic writing

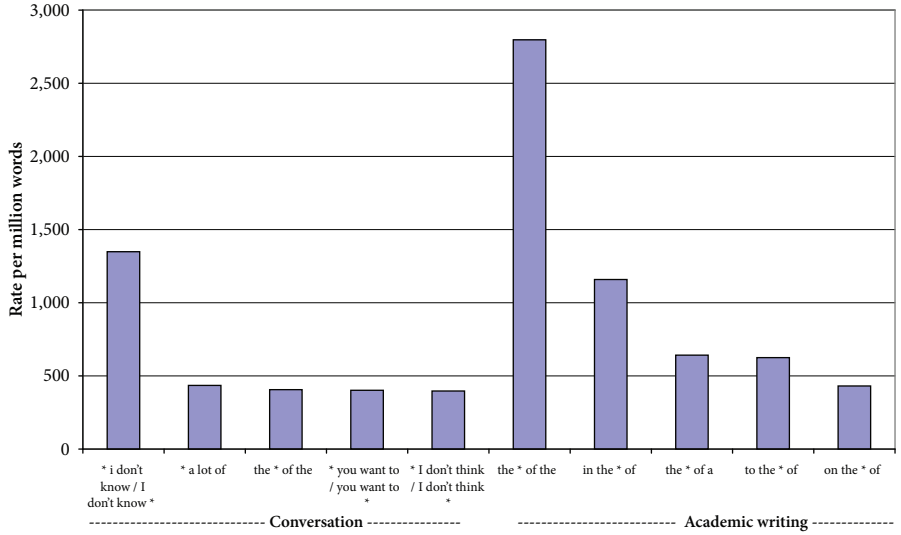

Figure 8. The most frequent patterns in conversation and academic writing

distribution, with academic writing employing some extremely high-frequency patterns. In particular, as Figure 8 shows, the pattern *the * of the* is more than twice as common as any individual pattern in conversation.

These high frequency patterns follow the same trends as described for Figure 5 above: 20 of the 22 most frequent patterns in conversation are continuous sequences with a preceding or following variable slot (pattern types *234 or 123*; see Table 3). In contrast, all nine of the most frequent patterns in academic writing are frames with an internal variable slot (pattern types 1*34 or 12*4; see Table 4).

Table 3. The most frequent patterns in conversation (all patterns occurring over 200 times per million words)

Pattern	Pattern type	Frequency per million	Most frequent multi-word sequence	% of pattern
* i don't know	*234	1,349	but i don't know	3
i don't know *	123*	1,349	i don't know if	12
* a lot of	*234	435	there's a lot of	4
the * of the	1*34	406	the end of the	10
* you want to	*234	402	do you want to	28
you want to *	123*	402	you want to do	11
i don't think *	123*	397	i don't think so	16
* i don't think	*234	397	but i don't think	6
* you have to	*234	387	do you have to	8
you have to *	123*	387	you have to do	10
* going to be	*234	328	it's going to be	11
going to be *	123*	328	going to be a	7
you know * i	12*4	325	you know what i	24
* you know what	*234	313	do you know what	11
you know what *	123*	313	you know what i	25
i'm going to *	123*	295	i'm going to go	8
* do you want	*234	275	what do you want	10
do you want *	123*	275	do you want to	42
i want to *	123*	250	i want to go	10
* i want to	*234	250	what i want to	6
don't want to *	123*	207	don't want to go	8
* don't want to	*234	207	i don't want to	50

Table 4. The most frequent patterns in academic writing (all patterns occurring over 200 times per million words)

Pattern	Pattern type	Frequency per million	Most frequent multi-word sequence	% of pattern
the * of the	1*34	2,797	the end of the	2
in the * of	12*4	1,159	in the case of	8
the * of a	1*34	642	the presence of a	3
to the * of	12*4	625	to the development of	2
on the * of	12*4	431	on the basis of	16
it is * to	12*4	369	it is possible to	16
at the * of	12*4	338	at the end of	19
it is * that	12*4	256	it is possible that	9
as a * of	12*4	253	as a result of	26
* part of the	*234	234	as part of the	7
one of the *	123*	234	one of the most	14
* one of the	*234	234	is one of the	12
* there is a	*234	221	that there is a	13
* a number of	*234	213	in a number of	9
the use of *	123*	204	the use of the	14

In most cases, the high frequency patterns in academic writing are extremely productive. For example, there are numerous fillers that occur in the highest frequency frame *the * of the*, and as Table 4 shows, the most common of these fillers (*end*) accounts for only 2% of the total occurrences of the pattern.

At the same time, these frames are distinctive in terms of the particular words that commonly serve as fillers. For example, four of the high-frequency patterns in academic writing are nearly identical, differing only in terms of the preposition that fills the first slot:

*in the * of*
*to the * of*
*on the * of*
*at the * of*

However, these frames select quite different sets of filler terms. The most distinctive frame is *at the * of*, which co-occurs most commonly with the fillers *end, time, beginning, level, expense, start, center/centre, top,* and *base*. None of these fillers commonly occurs in the other three frames.

At the other extreme, the frame *to the * of* is the least distinctive. Most of the fillers that commonly occur in this frame can also occur in the frames *in the * of* and *on the * of*, including the nouns *development, presence, use, study, question, number, nature, formation*.

The most common of these patterns — *in the * of* — takes several high frequency fillers that are distinctively used in this frame: *case, absence, form, context, course, process*. Similarly, the only high frequency filler for the pattern *on the * of* is distinctive for that frame: *basis*.

Taken together, the results of the present investigation indicate that the kinds of formulaic patterns common in conversation are fundamentally different from the formulaic patterns common in informational writing:

	Conversation	**Academic writing**
Predominant pattern	continuous sequence	frame with internal variable slot
Typical fillers in fixed slots	function words	function words
Typical fillers in variable slots	function words	content words

Conversation prefers fixed continuous sequences of words, with a preceding or following variable slot. Function words predominate in conversation, in both fixed and variable slots. In contrast, academic writing prefers formulaic frames with an internal variable slot. Function words predominate in the fixed slots, forming the 'frame', while content words predominate in the internal variable slot.

The analysis shows that any adequate methodological approach used for the study of longer formulaic sequences in English must account for the central role of function words, despite the high frequency of these words individually. Probabilistic methods, as they have been developed for the study of lexical collocations, are not well suited to this task. In addition, any omnibus statistic that measures the strength of association is not likely to capture the most important trends here: that these sequences are patterend in very different ways in terms of their fixed and variable slots.

A deeper exploration of content words in these patterns indicates that they are functionally, and perhaps grammatically, different in conversation versus academic writing. As noted above, most of the fillers of variable slots in the written patterns are content words. In most cases, these are nouns; there are usually dozens of nouns that can fill one of these slots, with no single noun accounting for a large proportion of the 4-word sequence. In contrast, there are relatively few content words that occur in conversational patterns. These words constitute a semi-closed class, half way in between function-word classes (pronouns, determiners, prepositions, etc.) and the fully open lexical classes of nouns, verbs,

adjectives, etc. In particular, the words *know, think (thought),* and *want* occur as content words in almost 50% of the conversational multi-word sequences that include a content word of any kind. The words *have, be, go, do,* and *like* (functioning as main verbs, not auxiliary verbs or semi-modals) account for most of the rest of the content words in these patterns. Thus we see a very different trend from academic writing: not only are content words generally dispreferred in conversational patterns, but when they do occur, they are usually drawn from a very restricted set of terms.

5. Conclusion

The present paper has explored the application of a radically corpus-driven approach to the study of formulaic language in English. Along the way, I have also considered the application of probabilistic methods for these research goals, suggesting that they are best suited for the study of collocations rather than extended (> two words) formulaic sequences.

The analyses here have revealed several surprising trends, most notably the conversational preference for continuous fixed sequences versus the written preference for formulaic frames with internal variable slots. The differing roles of content words in fixed versus variable slots was also uncovered by the analysis. These findings have important implications for our understanding of how discourse is produced and comprehended in speech versus writing, and important applied implications for how spoken versus written discourse is learned and taught.

The findings here are fully compatible with earlier corpus-based research that has argued that the grammar of speech is fundamentally different from the grammar of writing. In particular, clauses (including subordinate clauses) provide the foundation of discourse in speech, while phrases — especially noun phrases and prepositional phrases — provide the foundation of discourse in informational writing. Halliday has advocated this point of view since the 1970's, focusing especially on the importance of nominalization in academic written discourse (see, e.g., Halliday 1979; Halliday & Martin 1993; Halliday & Mathiessen 1999). Similarly, my own earlier studies of register variation have also repeatedly documented this pattern. For example, Dimension 1 in the original multi-dimensional study of English (Biber 1988: 104–108) showed that finite dependent clauses (*that*-clauses, WH-clauses, causative adverbial clauses, and conditional adverbial clauses) are characteristic of interpersonal spoken registers, in contrast with phrasal noun-modifying features (e.g. nouns, attributive adjectives, prepositional phrases), which are characteristic of formal written registers. A multi-dimensional study of discourse complexity (Biber 1992) confirms these patterns, while Biber (2006)

shows that similar discourse patterns distinguish university-level classroom teaching from written university registers.

In contrast to these earlier corpus-based studies, the present study employs a corpus-driven approach, based on only frequent combinations of simple word forms, but it identifies the same basic opposition between speech and writing. Formulaic language is very important in both conversational and written academic discourse, but it is realized in very different ways linguistically: fixed sequences that represent clause fragments in conversation, versus formulaic frames that consist of noun phrase and prepositional phrase fragments in academic writing.

On-going research is extending the analysis presented here to consider longer patterns, spanning over 5 words, 6 words, etc. Shorter combinations seem less interesting. Frequent two word combinations are mostly simple grammatical combinations, such as preposition+determiner or auxiliary verb+verb. Beyond that, two word combinations representing collocations have been studied extensively in previous research. In contrast, we know much less about the formulaic patterns represented by longer sequences of words. Thus, additional research is required to document the kinds of patterns found in longer sequences, and to develop a framework that captures the relationships among formulaic patterns of any length. In addition, all of these patterns and pattern types should be interpreted functionally, to describe their uses in spoken and written discourse contexts.

Finally, future research is planned to extend this line of inquiry to languages other than English, especially languages with different typological profiles. English has minimal inflectional morphology but a large set of grammatical function words. Those two factors are central determinants of the formulaic pattern types that are common in English. But how are formulaic expressions realized in other languages; for example, in morphology-rich languages like Finnish or Turkish? Or in isolating languages like Chinese? Obviously, different linguistic devices will be required to realize formulaic expressions in these languages, but it is not clear what those linguistic patterns will be. Further, it is not even clear that formulaic language will be equally important in all languages. This is an empirical question that can be best analyzed through corpus-driven analysis.

What does seem clear, though, is that the construct of formulaic language is complex and multi-faceted. The present paper has discussed two major parameters that are required for the description of formulaic language in English: collocations versus multi-word formulaic sequences, and fixed continuous sequences versus formulaic frames (with internal variable slots). The study has further shown that the nature of formulaic language in conversation is fundamentally different from formulaic language as it is realized in academic writing. It is reasonable to expect that these parameters of variation will need to be refined and extended in future research, as we consider additional spoken and written registers, and formulaic

expressions of differing types and lengths. It seems clear, though, that formulaic language is a complex construct that must be approached from several complementary perspectives for a full understanding.

Notes

* I would like to thank Viviana Cortes, Stefan Gries, Randi Reppen, and Norbert Schmitt for their helpful comments on earlier drafts of this paper.

1. Computer programs were developed by the author in Delphi Pascal for the analyses here. Similar kinds of analysis (of "phrase frames") can be carried out using *kfNgram*, an online software tool developed by William Fletcher (2007).

References

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, 101–122.

Altenberg, B. & Eeg-Olofsson, M. 1990. "Phraseology in spoken English". In J. Aarts & W. Meijs (Eds.), *Theory and Practice in Corpus Linguistics*. Amsterdam: Rodopi, 1–26.

Barlow, M. 2004. *Collocate 1.0*. Houston: Athelstan.

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

Biber, D. 1992. "On the complexity of discourse complexity: A multidimensional analysis". *Discourse Processes*, 15 (2), 133–163.

Biber, D. 2006. *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam/Philadelphia: John Benjamins.

Biber, D. & Barbieri, F. 2007. "Lexical bundles in university spoken and written registers". *English for Specific Purposes*, 26 (3), 263–86.

Biber, D., Conrad, S. & Reppen, R. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S. & Cortes, V. 2004. "If you look at…: Lexical bundles in university teaching and textbooks". *Applied Linguistics*, 25 (3), 371–405.

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

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

Cortes, V. 2004. "Lexical bundles in published and student disciplinary writing: Examples from history and biology". *English for Specific Purposes*, 23 (4), 397–423.

Ellis, N. C. 1996. "Sequencing in SLA: Phonological memory, chunking, and points of order". *Studies in Second Language Acquisition*, 18 (1), 91–126.

Ellis, N. C., Simpson-Vlach, R. & Maynard, C. 2008. "Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL". *TESOL Quarterly* 42 (3), 375–396.

Fletcher, W. 2007. *kfNgram*. Available at: http://www.kwicfinder.com/kfNgram/kfNgramHelp.html

Francis, G., Hunston, S. & Manning, E. (Eds.) 1996. *Collins COBUILD Grammar Patterns 1: Verbs*. London: HarperCollins.

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

Halliday, M. A. K. 1979. "Differences between spoken and written language: Some implications for literacy teaching". In G. Page, J. Elkins & B. O'Connor (Eds.), *Communication through Reading: Proceedings of the 4th Australian Reading Conference*. Adelaide: Australian Reading Association, 37–52.

Halliday, M. A. K. & Martin, J. R. 1993. *Writing Science: Literacy and Discursive Power* (Vol. first printed in 1993, reprinted in 1996). London: Falmer Press.

Halliday, M. A. K. & Matthiessen, C. 1999. *Construing Experience through Meaning: A Language-Based Approach to Cognition*. London: Cassell.

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

Howarth, P. 1998. "Phraseology and second language proficiency". *Applied Linguistics*, 19 (1), 24–44.

Hunston, S. & Francis, G. 2000. *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam/Philadelphia: John Benjamins.

Lindquist, H. & Mair, C. (Eds.) 2004. *Corpus Approaches to Grammaticalization in English*. Amsterdam/Philadelphia: John Benjamins.

Mahlberg, M. 2005. *English General Nouns: A Corpus Theoretical Approach*. Amsterdam/Philadelphia: John Benjamins.

McEnery T., Xiao, R. & Tono, Y. 2006. *Corpus-Based Language Studies*. London: Routledge

Moon, R. 1998. *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford: Clarendon Press.

Nesi, H. & Basturkmen, H. 2006. "Lexical bundles and discourse signaling in academic lectures". *International Journal of Corpus Linguistics* 11 (3), 283–304.

Oakes, M.P. 1998. *Statistics for Corpus Lingusitics*. Edinburgh: Edinburgh University Press.

Partington, A. 1998. *Patterns and Meanings*. Amsterdam/Philadelphia: John Benjamins.

Partington, A. & Morley, J. 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 am Main: Peter Lang, 179–192.

Renouf, A. & Sinclair, J.M. 1991. "Collocational frameworks in English". In K. Aijmer & B. Altenberg (Eds.), *English Corpus Linguistics*. London: Longman, 128–143.

Rohdenburg, G. & Mondorf, B. (Eds.) 2003. *Determinants of Grammatical Variation in English*. Berlin: Walter de Gruyter.

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

Schmitt, N., Grandage, S. & Adolphs, S. 2004. "Are corpus-derived recurrent clusters psycholinguistically valid?" In N. Schmitt (Ed.), *Formulaic Sequences*. Amsterdam/Philadelphia: John Benjamins, 127–152.

Sinclair, J. McH. 1991a. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Sinclair, J. McH. 1991b. "Shared knowledge". In *Georgetown University Round Table on Languages and Linguistics 1991*. Washington D.C.: Georgetown University Press, 489–500.
Sinclair, J. McH. 2001. Review of Biber et al., *The Longman Grammar of Spoken and Written English*. *International Journal of Corpus Linguistics* 6 (2), 339–359.
Stubbs, M. 1993. "British traditions in text analysis: From Firth to Sinclair". In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and Technology: In Honour of John Sinclair*. Amsterdam/Philadelphia: John Benjamins, 1–33.
Stubbs, M. 1995. "Collocations and semantic profiles: On the cause of the trouble with quantitative methods". *Functions of Language*, 2 (1), 1–33.
Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins.
Weinert, R. 1995. "The role of formulaic language in second language acquisition: A review". *Applied Linguistics*, 16 (2), 180–205.
Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.
Wray, A. & Perkins, M. 2000. "The functions of formulaic language: An integrated model". *Language and Communication*, 20 (1), 1–28.

Appendix I

Variability for the most common 4-word sequences in conversation, showing the extent to which each slot is variable or fixed. (100% = the slot is completely fixed in this sequence)

4-word sequence	Rate per million words	Slot 1 %	Slot 2 %	Slot 3 %	Slot 4 %	Pattern Type
i don't know if	162	89	92	93	12	123*
i don't know what	145	79	89	89	10	123*
do you want to	116	28	88	67	42	*23*
i don't want to	104	50	55	62	57	1234
you don't have to	86	53	56	57	57	1234
i was going to	85	51	74	44	82	12*4
you want me to	82	79	84	74	94	1234
or something like that	81	63	85	96	92	1234
you know what i	80	72	90	24	25	12**
i don't know how	80	78	91	86	5	123*
if you want to	78	19	77	59	54	*234
are you going to	68	59	63	66	71	1234
i don't think so	65	81	89	94	16	123*
i thought it was	51	62	32	45	77	1**4

4-word sequence	Rate per million words	Slot 1 %	Slot 2 %	Slot 3 %	Slot 4 %	Pattern Type
i don't know why	51	88	99	84	3	123*
to be able to	48	27	98	59	97	*234
but i don't know	48	3	89	78	36	*23*
i think it was	47	72	30	55	36	1*3*
going to have to	47	72	100	45	42	12**
you want to do	46	35	29	95	11	**3*
know what i mean	46	70	46	87	42	1*3*
what are you doing	46	50	92	87	32	123*
to go to the	44	34	60	55	23	*23*
do you want me	44	51	99	89	16	123*
what do you think	44	29	87	96	23	*23*
you don't want to	43	20	39	28	61	***4
and i don't know	43	3	86	81	38	*23*
the end of the	42	84	10	96	51	1*34
i'm not going to	42	27	60	67	86	*234
in the middle of	41	83	100	32	60	12*4
you want to go	41	31	40	99	10	**3*
it's going to be	38	11	52	100	53	*234
you have to do	38	25	25	94	10	**3*
all of a sudden	36	98	100	90	97	1234
do you know what	35	11	95	78	21	*23*
i don't think i	35	88	84	32	8	12**
i don't know where	35	79	91	89	2	123*
for a long time	35	36	100	96	86	*234
what do you mean	32	71	96	95	16	123*
do you have to	31	8	53	18	16	*2**
not going to be	31	9	66	100	20	*23*
at the same time	31	84	100	90	81	1234
i would like to	29	57	66	53	64	1234
put it in the	29	33	48	59	33	**3*
i mean i don't	29	93	30	71	15	1*3*
at the end of	29	35	99	39	57	*2*4
the rest of the	29	94	7	96	39	1*3*
what do you want	29	10	79	87	14	*23*

4-word sequence	Rate per million words	Slot 1 %	Slot 2 %	Slot 3 %	Slot 4 %	Pattern Type
i'm going to go	26	25	80	99	8	*23*
you have to be	26	21	33	99	6	**3*
want to go to	26	14	100	40	20	*2**
but i don't think	26	6	97	73	20	*23*
i thought you were	25	81	71	63	34	123*
would you like to	25	72	82	77	44	123*
going to be a	25	23	100	26	7	*2**
i want to go	25	19	23	100	10	**3*
thank you very much	25	82	99	75	93	1234
well i don't know	25	1	93	87	52	*234
you have to have	24	31	46	97	6	**3*
i'm going to have	24	21	94	100	8	*23*
i don't think it	23	83	75	23	5	12**
i mean you know	23	95	25	68	29	1*3*
the middle of the	23	94	5	92	46	1*3*
i don't think it's	22	84	74	68	5	123*
was going to say	22	43	80	100	13	*23*
i want to do	22	17	22	100	8	**3*
want to do it	22	11	99	21	17	*2**
know what it is	21	26	51	55	53	*234
i'm going to get	21	19	57	100	7	*23*
there's a lot of	21	4	100	53	80	*234
have to do it	20	10	91	16	13	*2**
i want to see	20	34	46	97	8	**3*
i'm going to be	20	6	71	100	7	*23*
i don't know that	19	71	33	20	1	1***
don't worry about it	19	59	75	97	64	1234
what did you say	18	46	71	42	23	*2**
don't know what it	18	46	83	27	9	*2**
i thought that was	18	67	49	16	46	1***
what did you do	18	50	41	51	23	1*3*
have to have a	18	19	92	24	24	*2**
don't want to go	18	14	60	100	8	*23*
i don't know whether	18	88	83	99	1	123*

4-word sequence	Rate per million words	Slot 1 %	Slot 2 %	Slot 3 %	Slot 4 %	Pattern Type
he was going to	17	10	60	38	75	*2*4
and then you can	17	55	62	49	13	12**
what do you do	17	45	39	56	9	**3*
i'm going to do	17	14	63	100	6	*23*
what i want to	16	6	45	37	54	***4
are we going to	16	61	15	62	67	1*34
you're not going to	16	10	73	48	81	*2*4
and you have to	16	4	30	41	48	****
going to have a	16	17	100	16	14	*2**
going to do it	16	8	98	18	13	*2**
you know i mean	16	70	56	96	8	123*
got a lot of	16	3	98	34	72	*2*4
know what to do	15	47	46	64	54	**34
how do you know	15	9	62	78	18	*23*
you can do it	15	32	30	15	20	****
i don't think you	15	84	77	29	3	12**
i want to get	15	22	16	98	6	**3*
you were going to	14	26	33	51	73	**34
you're going to be	14	4	70	98	14	*23*
we're going to have	14	12	91	98	14	*23*
and things like that	14	41	21	98	94	**34
i don't think he	14	85	87	77	3	123*
the other side of	14	98	56	68	43	123*
put it on the	13	27	45	26	22	****
you know if you	13	30	31	9	27	****
be able to get	13	64	71	100	7	123*
i tell you what	13	32	88	97	19	*23*
i mean if you	13	90	44	18	28	1***
going to get a	13	17	100	13	12	*2**
don't know what to	13	48	98	33	7	*2**
we're not going to	12	7	61	71	83	*234
i've got to go	12	32	97	98	19	*23*
go to the bathroom	12	60	93	100	9	123*
as long as you	12	91	30	100	21	1*3*

4-word sequence	Rate per million words	Slot 1 %	Slot 2 %	Slot 3 %	Slot 4 %	Pattern Type
what do you call	12	85	82	67	6	123*
to do with the	11	61	37	38	14	1***
if you're going to	11	11	41	76	78	**34
i think this is	11	58	35	20	47	1***
do you want some	11	36	100	64	4	*23*
do you want a	11	33	95	16	4	*2**
you're going to have	11	10	92	100	11	*23*
you don't need to	10	36	35	6	34	****
thought it was a	10	7	53	92	12	*23*
the back of the	10	86	2	75	33	1*3*
the top of the	10	39	2	86	45	**3*
nothing to do with	10	11	100	96	72	*234
that's going to be	10	2	52	100	45	*23*
the bottom of the	10	96	2	76	53	1*34
that sort of thing	10	60	40	98	45	1*3*
i'll tell you what	10	24	87	88	27	*23*
can i have a	10	5	72	43	27	*2**
oh i don't know	10	0	95	77	47	*23*
you might as well	10	39	78	100	100	*234
what did he say	10	60	86	21	52	12*4
no i don't think	10	2	100	89	22	*23*
that's what i mean	10	11	59	84	5	*23*
that's what i said	10	44	58	28	5	*2**
i said well i	10	53	50	16	14	12**

Appendix II

Variability for the most common 4-word sequences in academic prose, showing the extent to which each slot is variable or fixed. (100% = the slot is completely fixed in this sequence)

4-word sequence	Rate per million words	Slot 1 %	Slot 2 %	Slot 3 %	Slot 4 %	Pattern type
on the other hand	101	97	97	75	78	1234
in the case of	93	82	95	8	88	12*4

4-word sequence	Rate per million words	Slot 1 %	Slot 2 %	Slot 3 %	Slot 4 %	Pattern type
on the basis of	69	70	96	16	93	12*4
at the same time	69	91	96	86	79	1234
as a result of	67	83	87	26	71	12*4
the end of the	67	68	2	95	53	1*34
at the end of	66	52	94	19	92	12*4
it is possible to	61	98	84	16	61	12*4
the nature of the	53	65	1	97	51	1*34
in the absence of	48	57	98	4	95	12*4
at the time of	47	71	92	14	67	12*4
in the form of	45	55	90	3	75	12*4
in the presence of	41	24	98	3	87	*2*4
the extent to which	40	95	55	99	97	1234
the way in which	40	72	33	96	96	1*34
it is necessary to	38	80	83	10	75	12*4
per cent of the	37	98	94	68	46	123*
the fact that the	34	90	27	94	21	1*3*
one of the most	33	64	98	91	14	123*
to be able to	32	30	97	22	90	*2*4
can be used to	32	42	97	24	41	*2**
is likely to be	31	31	21	100	45	**3*
it is important to	31	88	88	8	58	12*4
the use of the	30	42	1	94	14	**3*
that there is a	30	13	41	75	33	**3*
in the same way	29	73	99	69	31	123*
in terms of the	29	48	18	95	26	**3*
on the one hand	28	96	95	21	91	12*4
the rest of the	28	95	1	88	70	1*34
is one of the	28	12	37	93	72	**34
the case of the	26	85	0	62	23	1*3*
in the context of	25	72	90	2	87	12*4
as we have seen	25	57	96	98	64	1234
it is possible that	24	97	87	9	24	12**
it is difficult to	24	67	73	6	87	12*4
at the beginning of	23	55	95	6	79	12*4

4-word sequence	Rate per million words	Slot 1 %	Slot 2 %	Slot 3 %	Slot 4 %	Pattern type
it is clear that	23	98	70	9	76	12*4
the size of the	22	67	0	97	60	1*34
the role of the	21	72	0	95	28	1*3*
the development of the	21	61	0	96	15	1*3*
the basis of the	21	86	0	79	21	1*3*
are likely to be	21	21	26	98	44	**3*
in a number of	20	9	52	15	95	*2*4
it should be noted	20	84	69	94	31	123*
in the development of	19	13	90	1	92	*2*4
in the course of	19	50	94	1	92	12*4
the presence of a	19	89	3	100	11	1*3*
in the present study	18	60	97	66	49	123*
in the process of	18	24	88	1	62	*2*4
as part of the	18	7	30	98	40	**3*
an increase in the	18	38	53	98	34	*23*
the time of the	18	77	0	66	27	1*3*
the division of labour	18	57	62	91	66	1234
the surface of the	18	57	0	89	62	1*34
should be noted that	18	59	94	51	74	1234
has been shown to	18	50	99	20	45	12**
as well as the	17	57	32	75	14	1*3*
in the number of	17	8	44	1	89	***4
in such a way	17	94	94	96	45	123*
the structure of the	17	46	0	85	44	**3*
the shape of the	17	76	0	97	71	1*34
the secretary of state	17	81	59	98	90	1234
the results of the	17	80	0	68	24	1*3*
the position of the	17	59	0	89	53	1*34
that there is no	17	10	92	56	19	*23*
has been suggested that	17	83	97	28	70	12*4
for the first time	16	82	98	89	49	123*
by the fact that	16	10	97	50	94	*234
as a function of	16	43	96	6	98	*2*4

4-word sequence	Rate per million words	Slot 1 %	Slot 2 %	Slot 3 %	Slot 4 %	Pattern type
the relationship between the	16	58	16	78	31	1*3*
the beginning of the	16	92	0	92	40	1*3*
to ensure that the	16	76	17	97	40	1*3*
it has been shown	16	39	98	97	14	*23*
it can be seen	16	31	68	98	17	*23*
to the development of	15	10	88	2	96	*2*4
at the level of	15	24	82	4	90	*2*4
the ways in which	15	40	12	92	95	**34
the base of the	15	84	0	83	76	1*34
is based on the	15	24	30	85	40	**3*
the value of the	14	47	0	80	21	**3*
the top of the	14	80	0	88	69	1*34
more likely to be	14	15	75	100	37	*23*
to do with the	13	91	19	64	43	1*3*
in addition to the	11	47	18	47	37	****
with respect to the	11	64	52	96	30	123*
the sum of the	11	87	0	92	48	1*3*
the magnitude of the	11	87	0	98	58	1*34
are more likely to	11	28	44	54	92	**34
should be able to	9	9	94	13	85	*2*4
is the same as	9	25	98	88	42	*23*
is equal to the	9	40	6	98	55	**34
as shown in figure	8	22	65	88	21	*23*
for the most part	6	85	90	65	62	1234

Author's address

Douglas Biber
Northern Arizona University
English Department
NAU
Flagstaff, AZ 86011-6032
USA

Douglas.Biber@NAU.EDU