Register as a predictor of linguistic variation

DOUGLAS BIBER

Abstract

Over the last two decades, corpus analysis has been used as the basis for several important reference grammars and dictionaries of English. While these reference works have made major contributions to our understanding of English lexis and grammar, most of them share a major limitation: the failure to consider register differences. Instead, most reference works describe lexico-grammatical patterns as if they applied generally to English.

The main goal of the present paper is to challenge this practice and the underlying assumption that the patterns of lexical-grammatical use in English can be described in general|global terms. Specifically, I argue that descriptions of the average patterns of use in a general corpus do not accurately describe any register. Rather, the patterns of use in speech are dramatically different from the patterns in writing (especially academic writing), and so minimally an adequate description must recognize the two major poles in this continuum (i.e., conversation versus informational written prose).

The paper begins by comparing two general corpus approaches to the study of language use: variationist and text-linguistic. Although both approaches can be used to investigate the use of words, grammatical features, and registers, the two approaches differ in their bases: the first gives primacy to each linguistic token, while the second gives primacy to each text. This difference has important consequences for the overall research design, the kinds of variables that can be measured, the statistical techniques that can be applied, and the particular research questions that can be asked. As a result, the importance of register has been more apparent in text-linguistic studies than in studies of linguistic variation.

The bulk of the paper, then, argues for the importance of register at all linguistic levels: lexical, grammatical, and lexico-grammatical. Analyses comparing conversation and academic writing are discussed for each level, showing how a general 'average' description includes some characteristics that are not applicable to one or the other register, while also omitting other important patterns of use found in particular registers.

Corpus Linguistics and Linguistic Theory 8–1 (2012), 9–37 1613-7027/12/0008–0009
DOI 10.1515/cllt-2012-0002

Keywords: *register differences, linguistic variation, conversation, academic writing, research designs*

1. Introduction

One major contribution of corpus research over the past 40 years has been the increasing awareness that lexis and grammar are intimately intertwined. Numerous studies have exploited corpus resources to describe the systematic lexical associations of a target grammatical construction (cf. the survey of studies in Kennedy 1998: 121–154). For example, some of the most common verbs in English occur most of the time in the simple present tense (e.g., *think, know, want, mean*), while others occur more often in the simple past tense (e.g., *said, came, took*) (Kennedy 1998: 123). The modal *would* usually occurs with a simple verb in written academic writing (*the expense would fall*), while the modal *can* more often occurs with *be* + past participle (*the procedures can be applied to . . .*) (Kennedy 1998: 133). The preposition *between* most commonly occurs as a noun modifier in written English (e.g., *difference between, relationship between, agreement between*), in contrast to the preposition *through*, which more often has an adverbial function (e.g., *go through, pass through, come through*) (Kennedy 1998: 142–143; cf. Kennedy 1991). A more recent study of this type is Römer's (2005) detailed description of the verbs associated with progressive aspect.

Several major reference grammars have also employed corpus investigations to identify the words associated with grammatical constructions (e.g., lists of the verbs and adjectives that can control a *that*-clause or a *to*-clause). One of the earliest grammars to include extensive lexical information of this type is the *Comprehensive Grammar of the English Language* (Quirk et al. 1985; see also Quirk et al. 1972), while the *Collins COBUILD English Grammar* (1990), the *Longman Grammar of Spoken and Written English* (Biber et al. 1999), and the *Cambridge Grammar of English* (Carter and McCarthy 2006) are more recent examples.

These grammars all take a deductive ('corpus-based') approach: grammatical constructions are distinguished on the basis of traditional linguistic criteria, and then the set of words associated with those constructions are identified through corpus analysis. In contrast, the 'pattern grammar' reference books (Francis et al. 1996, 1998) take the opposite approach, beginning with words and then identifying the (grammatical) "phraseology frequently associated with (a sense of) a word" (Hunston and Francis 2000: 3). These books show that there are systematic regularities in the associations between sets of words, grammatical frames, and particular meanings on a much larger scale than it

could have been possible to anticipate before the introduction of large-scale corpus analysis.

Surprisingly, most of these previous reference books and studies do not report quantitative findings, and it is not clear that specific quantitative analyses were undertaken as part of the research endeavor. These studies sometimes suggest that they are based on quantitative analysis, claiming to describe the words that 'commonly', 'frequently', or 'usually' occur with a given construction. However, in most cases, no actual quantitative findings are reported.[1]

There are major advantages to reporting quantitative findings. First, a quantitative approach requires explicit operational definitions and accountability in the linguistic analysis. When specific quantitative findings are not reported, the reader must simply take it on faith that the identified lexical-grammatical patterns are in fact 'common' or 'frequent'. Similar to observational research in the natural sciences and social sciences, corpus research is based on 'samples' (corpora) which are carefully designed to represent 'populations' (particular registers and dialects in a particular language). However, it is virtually inconceivable in other disciplines that a researcher would go to the trouble of designing and collecting such samples but then not analyzing the quantitative distributions of variables in those samples.

But more importantly, quantitative findings provide essential information about language use by documenting the extent to which a lexical-grammatical association holds. That is, most linguistic phenomena are not distributed in a simple binary opposition of 'frequent' versus 'rare.' Rather, there is normally a continuous range of variation, and quantitative findings are required to adequately describe those patterns. As a result, all the analyses presented below employ quantitative analysis.

However, there is a second characteristic of most major reference works that is less often recognized as a shortcoming: the failure to consider register differences (cf. Biber and Conrad 2009). That is, most reference works describe lexico-grammatical patterns as if they applied generally to English. When quantitative findings are reported, they are based on an entire corpus (e.g., the Brown Corpus or the ICE-GB Corpus), which is taken to represent English generally.

The main goal of the present paper is to challenge this practice and the underlying assumption that the patterns of lexical-grammatical use in English can be described in general/global terms. Specifically, I argue that descriptions of the average patterns of use in a general corpus do not accurately describe any register. Rather, the patterns of use in speech are dramatically different from the patterns in writing (especially academic writing), and so minimally an adequate description must recognize the two major poles in this continuum (i.e., conversation versus informational written prose).

The paper begins by comparing two general corpus approaches to the study of language use: variationist and text-linguistic. Although both approaches can

be used to investigate the use of words, grammatical features, and registers, the two approaches differ in their bases: the first gives primacy to each linguistic token, while the second gives primacy to each text. This difference has important consequences for the overall research design, the kinds of variables that can be measured, the statistical techniques that can be applied, and the particular research questions that can be asked. As a result, the importance of register has been more apparent in text-linguistic studies than in studies of linguistic variation.

The bulk of the paper, then, argues for the importance of register at all linguistic levels: lexical, grammatical, and lexico-grammatical. Analyses comparing conversation and academic writing are discussed for each level, showing how a general 'average' description includes some characteristics that are not applicable to one or the other register, while also omitting other important patterns of use found in particular registers. (The data for several of these case studies are taken from the *Longman Grammar of Spoken and Written English*, referred to as *LGSWE* below.)

2. Perspectives on 'frequency': Linguistic versus text-linguistic variation

Corpus-based studies generally have one of two primary research goals: 1) to describe the variants and use of a word or linguistic structure, or 2) to describe differences among texts and text varieties, such as registers or dialects. The first goal relates to classic studies of linguistic variation (e.g., the choice between active and passive voice), while the second is used to investigate text-linguistic variation, contrasting the words and grammatical structures typically found in different texts and varieties.

Biber, Conrad and Reppen (1998: 269–274) as well as Biber and Jones (2009) identify three major types of research design that have been employed in corpus research. The primary difference among these research design types is the unit of analysis (or the 'observations'), which in turn makes each design type appropriate for one of the above two research goals:

- Type A studies: the unit of analysis is each occurrence of a linguistic feature. Type A studies are thus designed for Research Goal 1 (describing the variants of a linguistic structure).
- Type B studies: the unit of analysis is each individual text. Type B studies are thus designed for Research Goal 2 (describing the differences among texts and text varieties).
- Type C studies: the unit of analysis is the entire corpus (or different subcorpora). Type C studies can be used for either Research Goal 1 or 2, but they do not permit the use of inferential statistics.

The units of analysis are the 'observations' that are described in a study. For the most part, the observations in Type A studies do not have quantitative characteristics, while the observations in Type B studies are analyzed in terms of quantitative characteristics. (Type C studies differ from both of the others in that there are actually very few observations – usually only 2 or 3 observations – because each sub-corpus is treated as an observation.)

For example, a Type A study of relative clauses might have the goal of predicting the choice of relative pronoun (*who*, *which*, *that*). Each relative clause would be an observation, coded for restrictive versus non-restrictive function and for the animacy of the head noun. All three variables (relative pronoun, clause type, head noun type) in this study would be nominal rather than numeric.[2]

In this case, descriptive statistics give the frequencies for each combination of categories (e.g., how many occurrences of the relative pronoun *that* are used with animate head nouns). However, it is difficult to document variation (or dispersion) across the corpus, and the distribution of linguistic variants across texts is not considered in this type of analysis (see also Gries 2006). Most Type A studies obtain frequencies for the corpus as a whole, but give no consideration to variation among the texts in a corpus.

Statistically, this type of design must be analyzed using non-parametric techniques, such as chi-squared or log-likelihood. As a result, a Type A study is ideal for studying the proportional preference for one or another variant, or the proportional extent to which a linguistic variant occurs with particular contextual factors. However, this design type is not well suited to studying the overall extent to which a linguistic feature is used in texts.

The important point here is that Type A research designs do not provide the basis for determining rates of occurrence, so they cannot be used to determine if a feature or variant occurs more often in one register or another. This is potentially confusing, and even published research studies sometimes make this mistake in interpreting statistical analyses: describing proportional preference for one variant over another as if it was the same as a higher rate of occurrence.

Type A studies can tell us what the preferred variant is in a register, and how registers differ in their reliance on a particular variant. For example, Figure 1 shows the proportional use of *that* versus 0 complementizer in verb + *that*-clause constructions, as in:

The commission agreed that this solution . . .
versus
I thought [0] you did it

Figure 1 compares conversation, newspaper writing, and academic prose, based on a sample of 1,000 *that*-clauses taken from each of the registers. (The actual proportions are given in Table 1; cf. LGSWE 1999: 680).

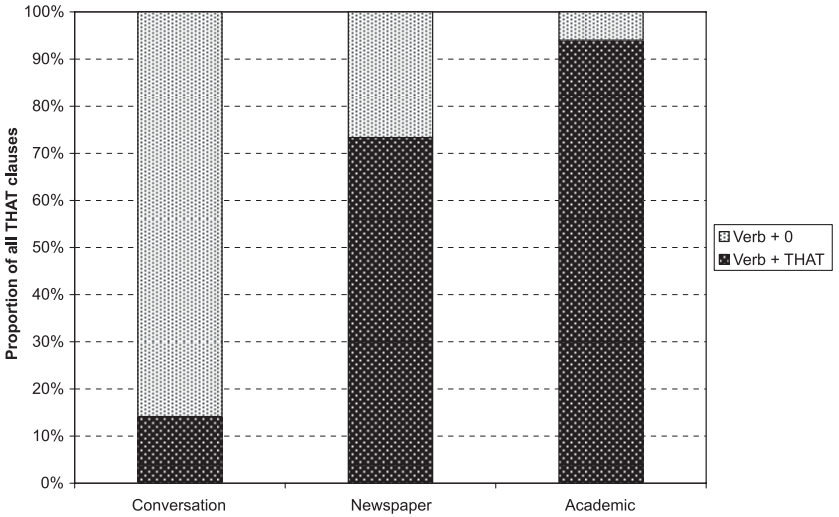

Figure 1. *Proportional preference for* Verb + that *versus* Verb + 0 *in three registers*

Table 1. *Proportional preference for* Verb + that *versus* Verb + 0*, based on a sample of 1,000 that-clauses taken from conversation, newspaper writing, and academic prose*

	Conversation	Newspaper	Academic prose
Verb + *that*	141 (14%)	733 (73%)	940 (94%)
Verb + 0	859 (86%)	267 (27%)	60 (6%)
Total	1,000 (100%)	1,000 (100%)	1,000 (100%)

It would be easy to look at Figure 1 (and Table 1) and incorrectly conclude that verb + *that* is most common in academic writing: 6 times more frequent than in conversation, and also considerably more frequent than in newspaper writing. But in fact, Table 1 does not provide the basis for such conclusions, and they are actually incorrect.[3]

Rather, Figure 1 (and Table 1) shows that when a *that*-clause is used in academic prose, it will almost always retain the *that* complementizer. When a *that*-clause is used in conversation, it will usually omit the complementizer. *That*-clauses in newspaper writing usually retain the complementizer, but this preference is less pronounced than in academic prose. These are genuine register differences. However, it would be incorrect to therefore conclude that *that*-retention is more common in academic prose than in newspaper writing or conversation.

In contrast, Figure 2 presents that the actual rates of occurrence for verb + *that*. Figure 2 shows that the rate of *that*-retention is much higher in newspaper

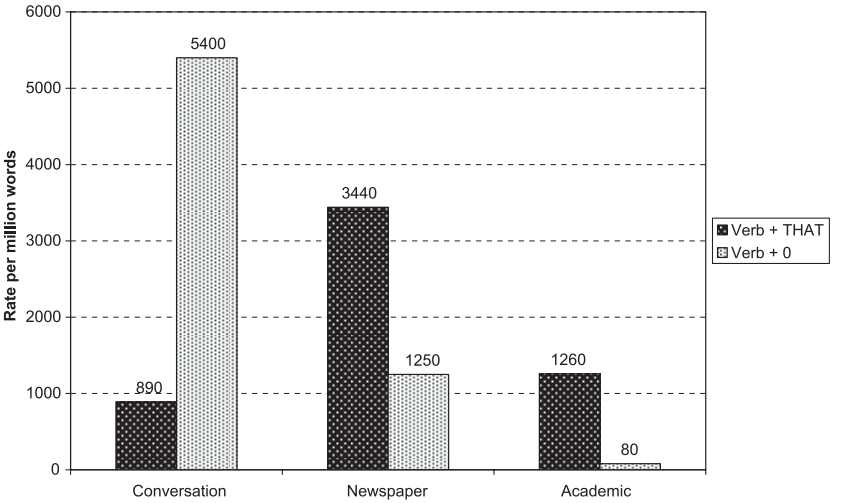

Figure 2. *Rates of occurrence for* Verb + that *and* Verb + 0 *in three registers*

writing than in academic writing, and the rate in conversation is only somewhat lower than in academic writing. This is because *that*-clauses overall are much more frequent in conversation and newspaper writing than in academic writing. As a result, both variants (with *that* and 0) occur with much higher rates in newspaper writing than in academic writing.

It is surprisingly common for researchers to confuse these two perspectives on variation, or to at least use statements that are misleading to the naïve reader. The main problem here has to do with claims that a linguistic feature is "frequent." Consider, for example, the following statement from Szmrecsanyi and Hinrichs (2008: 297): "The *s*-genitive is, on the whole, more frequent in spoken data than in written . . ." It would be natural to interpret this statement to mean that a speaker will produce more *s*-genitives than a writer will. Or put another way, a listener will encounter more *s*-genitives in a conversation than a reader would in a written text. However, that interpretation is not intended by Szmrecsanyi and Hinrichs, and in fact, it is not accurate.

The pattern being described by Szmrecsanyi and Hinrichs is one of proportional preference, not frequency of occurrence in texts. That is, *s*-genitives are proportionally preferred over *of*-genitives in speech, while *of*-genitives are proportionally preferred over *s*-genitives in writing: "FRED [i.e. a spoken corpus] exhibits the highest percentage of the *s*-genitive (59.6%), Brown [i.e., a written corpus] (36.2%) the lowest." (Szmrecsanyi and Hinrichs 2008: 297). But, from a text-linguistic perspective, *s*-genitives are actually much more frequent in writing than in speech. Thus, Figure 4.6 in LGSWE (1999: 302) shows

that there are only c. 800 *s*-genitives per million words of conversation, in contrast to c. 2,300 *s*-genitives per million words in academic writing, and c. 9,000 *s*-genitives per million words in newspaper writing.

The pattern here is the same as the case study for *that*-retention presented above. Thus, when a speaker uses a genitive construction, it is most likely to be an *s*-genitive. So, considering only conversation, *s*-genitives are more frequent than *of*-genitives. However, genitives overall are much more frequent in writing than in speech. As a result, when speech is compared to writing, both *of*-genitives and *s*-genitives occur more frequently in writing. That is, even though the *s*-genitive is proportionally preferred in conversation, it still is much less frequent than the *s*-genitive in writing.

The point here is not to criticize the Szmrecsanyi and Hinrichs (2008) study, which is an exemplary study employing a carefully considered research design and sophisticated statistical methods to analyze this aspect of grammatical variation. Rather, the point is to emphasize how easy it is to confound the variationist and the text-linguistic perspectives when reporting frequency results. This is not an obscure methodological quibble. Rather, the two perspectives are completely different in their practical implications. The variationist perspective has the goal of comparing linguistic variants: whether one or the other variant is preferred. These preferences can be compared across registers, but that analysis cannot tell us the actual extent to which a variant is used in texts. In contrast, the text-linguistic perspective (a Type B study) has the goal of providing a linguistic description of texts, by describing the density of grammatical features in texts. These studies directly tell us the density of occurrence of a feature (or variant) in texts from different registers.

There are a few general points worth emphasizing here. The first is the importance of distinguishing between variationist research designs and text-linguistic designs: variationist designs investigate proportional preferences, while text-linguistic designs investigate the rates of occurrence in texts. But a more general point is a cautionary one: the text-linguistic perspective is often the more natural interpretation, and thus it is easy for authors (and readers) to misleadingly use the language of 'frequency'. When a linguistic feature is described as occurring 'frequently', we expect to encounter numerous occurrences of the feature in texts (the text-linguistic perspective). The proportional perspective is more difficult to describe: that when a linguistic feature does occur, it usually has certain characteristics – even if the feature itself occurs infrequently. Thus, it is essential to be explicit about the nature of the patterns in variationist studies: that they represent proportional preferences but not necessarily frequent occurrence in texts.

In summary, a Type A research design – studies of linguistic variation where each linguistic token is an observation – cannot describe the rates of occurrence in texts and registers. This design type can identify register influences on

linguistic variation – different proportional preferences in different registers – but it cannot tell us how frequently we will encounter a linguistic feature in texts from a register. In contrast, a Type B study – studies of text-linguistic variation where each text is an observation – is specifically designed for such research questions. However, as the following sections show, register differences are important in both design types.[4]

3. Lexical patterns and register

3.1 *Lexis from the perspective of linguistic variation: Collocational studies*

The importance of register for vocabulary is well-established from corpus research (see, e.g., Kennedy 1998: 97–100). To take an easy example, the pronouns *I* and *you* are among the most frequent words in the spoken London-Lund Corpus, but considerably less frequent in the written Brown Corpus. ELT dictionaries like the *Longman Dictionary of Contemporary English* (2009) provide detailed information of this type, explicitly identifying the most frequent words in speech versus the most frequent words in writing.

Despite this high level of awareness concerning the importance of register differences for word frequencies, most studies of collocation and extended lexical patterns have disregarded register. The unstated assumption has been that the lexical collocations of a word will remain constant, regardless of register.

Many studies have investigated the preferred collocates of specific target words. For example, Sinclair (1991) describes the uses of phrasal verbs with *set* and collocates of the word *back*. Hunston (2002: Chs. 3 and 4) discusses the phraseological patterns of several target words, such as *recipe, initiative, condemn, suggestion, point, gaze, leak,* and *shoulder.* Partington (1998), in a book-length treatment of collocation, provides detailed descriptions of the phraseological patterns for *sheer, pure, complete, absolute, correct, absolutely, completely, entirely.* These studies provide detailed descriptions of the collocations and preferred uses of a specific target word, and further illustrate how supposed synonyms are not in fact identical in meaning or use when considered from this perspective. However, these studies are typical in that they include no mention of register or the possibility of different word uses in different registers.

However, corpus investigation of common words shows that this disregard for register is not justified. For example, contrast the common content-word collocates for the verbs *have, make,* and *take* in conversation versus informational writing (taken from Conrad and Biber 2009: 13–20):

have
Conversation:
dinner, lunch, a drink, fun, a good time, trouble, a hard time, a|no problem with, kids, children, a baby, a|the chance, an|no idea, a question
Informational writing:
an|no|little effect|impact|influence on, the advantage of, a range of, a wide variety of, little|no evidence of, no knowledge of, the potential for, little sympathy for, implications for, an interest in, a role in

make
Conversation:
the bed, a phone call, a joke, (a) noise, a sound, an appointment, a deal, plans to, a living, money, a difference, (a) decision(s), an effort, a mistake, (no) sense, fun of, time for, sure
Informational writing:
assumptions, comparisons, judgments, choices, decisions, predictions, recommendations, (no) sense, use of, reference to

take
Conversation:
a photo|picture, a bath|shower, a nap, a break, it easy, place, a minute, time, classes, a test, a message, notes, a car, the bus|train, a ride, a right|left [turn], a look at, care of, charge of, responsibility for, advantage of, forever, turns
Informational writing:
action, the initiative, the lead in, steps to, the position that, the view that, account of, into account, part in, advantage of, precedence over, the form of, the shape of

From a text-linguistic perspective, the verbs *have, make,* and *take* are considerably more frequent in conversation than in informational writing. However, the perspective of linguistic variation asks different questions: When these verbs are used in conversation, what are the most common collocates? When these verbs are used in informational writing, what are the most common collocates? Are the preferred collocates in conversation the same as those in informational writing?[5]

The answers to these questions indicate that register is a fundamentally important organizing factor for studies of collocation. All three verbs have strong collocational associations in both conversation and informational writing. However, those associations are almost entirely non-overlapping.

A collocational analysis of these verbs in a general multi-register corpus (cf. BNC) might identify many of these combinations. But such an analysis would miss the point that these are not general collocations. In informational writing, it is rare to find uses like *have lunch, have fun, make a phone call, make a deal,*

take a break, take care of, etc. Similarly, in conversation, it is rare to find uses like *have implications for, make assumptions about, take precedence over*. These are all strong collocations, but they are tied to a particular domain of use, and thus an essential component of their analysis should be documentation of the register where they typically occur.

3.2 *Text-linguistic lexical variation: Frequent lexical sequences*

Text-linguistic studies ask a different type of research question from collocational studies: What are the most common lexical sequences found in texts (or registers)? That is, in a collocational study of linguistic variation, particular target words are selected for investigation, and then the analysis identifies the preferred collocates for each word. There is little consideration given to whether the target word is rare or frequent in texts. Rather, the focus is on the preferred collocates when the target word occurs, regardless of how often the word occurs in texts.

In contrast, in a text-linguistic study, no words are pre-selected for investigation. Rather, the goal is to analyze the texts themselves, to identify the most frequent lexical sequences in those texts. Given this goal, it is natural that text-linguistic studies of lexical sequences would consider register differences; and as a result, these studies have consistently identified fundamentally different kinds of lexical patterns in different registers.

For example, one series of such studies have been carried out under the rubric of 'lexical bundles' (e.g., Biber et al. 1999: Ch. 13; Biber et al. 2004). Lexical bundles are identified using a frequency-driven approach. In the initial study of English lexical bundles (Biber et al. 1999), 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 carried out by Biber et al. (2004). 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: Ch. 13) compared the patterns of use in conversation and academic prose, based on

analysis of c. 5-million-word sub-corpora from each register. 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.

The most important finding from these studies for the purposes of the present paper is that the set of common lexical bundles in conversation is completely different from the set of common lexical bundles in academic writing. For example, six 4-word bundles are extremely frequent in conversation, occurring over 100 times per million words: *I don't know if, I don't know what, do you want to, I don't want to, I was going to, are you going to*. In contrast, only two 4-word bundles occur over 100 times per million words in academic writing: *on the other hand, in the case of*. Other frequent 4-word bundles in academic writing include: *one of the most, the nature of the, as a result of, on the basis of, in the form of*.

In sum, register is fundamentally important for the description of frequent lexical sequences, to a much greater extent than previously anticipated. In fact, the sets of common lexical bundles are nearly disjunctive between conversation and academic prose. For example, only four of the 133 four-word lexical bundles that occur over 20 times per million words in conversation were also found to occur over 20 times per million words in academic writing – an overlap of only three percent. All other lexical bundles were found to be distinctive for conversation versus academic prose.

The lexical bundle framework was extended in Biber (2009b) to allow for variable patterns, identifying differences between registers at an even more basic level. In that study, each 4-word lexical bundle was coded for its 'pattern type', depending on the extent to which each slot was variable or fixed. 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. For example, the sequence ____ *the case of* occurred 617 times in the academic writing corpus analyzed for this study. 506 of those occurrences – or 82% – were preceded by the word *in*. Thus, *in* is a 'fixed' slot in this 4-word sequence. In contrast, the sequence *in the* ____ *of* occurred 6,325 times in this corpus, and the most common filler – *case* – occurred only 506 times (or 8% of the total). Thus, the third word is a variable rather than fixed slot in this sequence.

The 2009 study shows that spoken discourse relies heavily on 3-word and 4-word fixed sequences of words, like *I don't know *, * you want to, I don't want to,* and *are you going to.* In contrast, written academic discourse is composed of lexical patterns that consist of invariable function words with an

intervening variable slot filled by many different possible content words (e.g., *the * of the, in the * of, to the * of*).

These different patterns also have strong grammatical correlates. For example, the continuous fixed sequences in conversation consist of both function words and content words. In contrast, the fixed slots in the academic writing patterns are usually function words, while the variable slots are usually content words. However, there are other differences having to do with the structural correlates of these lexical patterns: verb phrase and clause fragments in the case of conversational lexical sequences, but noun phrase and prepositional phrase fragments in the case of the academic writing patterns. I return to these associations in Section 5 below.

4. Grammatical variation and register

4.1 *Grammar from the perspective of linguistic variation*

Section 2 above has already introduced studies of grammatical variation that treat each token of the grammatical feature as an observation. For example, Figure 1 (above) showed how registers can differ in their proportional preference for a grammatical variant, while a comparison of Figure 1 and Figure 2 showed how the rates of occurrence for variants in texts provide a dramatically different perspective on register differences.

In recent years, grammatical variation has been investigated using sophisticated multivariate statistical techniques, such as logistic regression (e.g., Szmrecsanyi 2005; Hinrichs and Szmrecsanyi 2007). In general, these studies have not considered register as a predictor. However, when the influence of register/genre has been investigated, as in Riordan (2007) and Szmrecsanyi and Hinrichs (2008), it has been found to be an important predictor; for example, ". . . we have seen that more often than not, individual factors – for instance, possessor animacy or thematicity of the possessor – have fairly different impacts in spoken and written data." (Szmrecsanyi and Hinrichs 2008: 307)

The advantage of multivariate techniques like logistic regression is that they are able to isolate the predictive strength of a factor when numerous other factors are also considered. However, the importance of register for studies of grammatical variation can also be illustrated with simple descriptive statistics.

For example, we can further explore the retention versus omission of *that* in *verb + that/0* constructions (introduced in Section 2 above). Several grammatical factors have been hypothesized to influence this choice, such as the frequency of the matrix verb, and the presence of an intervening NP between

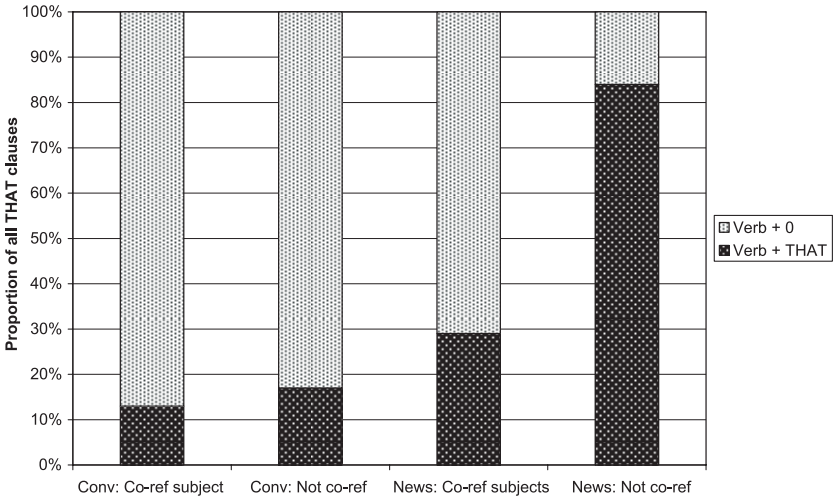

Figure 3. *Proportional preference for* Verb + that *versus* Verb + 0*: The influence of co-referential subjects, in conversation vs. newspaper prose*

the matrix verb and the *that*-clause. One factor that has been hypothesized to favor *that*-omission is co-referential subjects in the matrix clause and the *that*-clause, as in:

I thought [0] I would go

In this example, it is not grammatically possible to interpret the second *I* as the direct object as *thought*. Thus, the use of a co-referential subject in the *that*-clause is a relatively explicit signal that a new clause is beginning, even without the complementizer *that*.

In contrast, when the two subjects are not co-referential, there is greater need for the complementizer *that* to signal the start of a *that*-clause, as in:

The Secretary argued that armed intervention was not the answer.

This factor is interesting for our purposes here because its influence is dramatically different in different registers. Thus, Figure 3 above (based on Biber et al. 1999: 681) shows that the choice of grammatical subject has little influence in conversation: the complementizer *that* is omitted over 80% of the time, regardless of whether the grammatical subjects are co-referential or not. However, we see a dramatically different pattern in newspaper writing: when the construction has co-referential subjects, the complementizer *that* is omitted over 70% of the time, while *that* is omitted only c. 15% of the time when

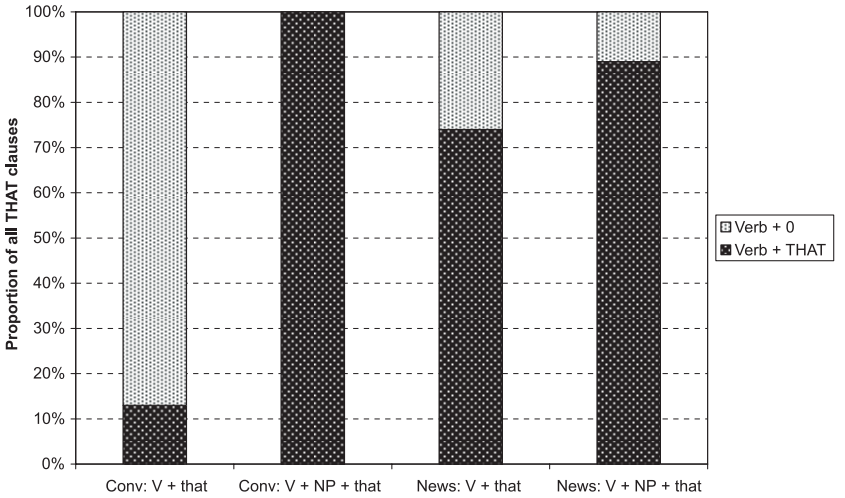

Figure 4. *Proportional preference for* Verb + that *versus* Verb + 0*: The influence of an intervening NP, in conversation vs. newspaper prose*

the subjects are not co-referential. Thus, the strength of this factor is mediated by register: essentially no influence in conversation (because *that*-omission is already the norm), versus a very strong influence in newspaper writing (because *that*-retention is the general register norm).

Other grammatical factors favor the retention of *that*, such as the presence of a noun phrase intervening between the matrix verb and *that*-clause (e.g., *They told him that it's dangerous*). These grammatical factors show the opposite interaction with register, having a very strong influence in conversation but little influence in informational writing. Thus, Figure 4 above (based on Biber et al. 1999: 682) shows that the presence of an intervening noun phrase in conversation results in the complementizer *that* almost always being retained, while it is retained less than 15% of the time when there is no intervening noun phrase. In contrast, the presence of an intervening noun phrase has little influence in newspaper writing, because the norm is already to retain the complementizer *that*. Thus, even without an intervening noun phrase, the complementizer *that* is retained c. 75% of the time.

In sum, similar to the collocational patterns for individual words, register is a strong predicting factor for studies of grammatical variation. This factor interacts with other contextual influences: a contextual factor with a strong influence in one register might have a minimal influence in another register. As the following section shows, grammatical differences across registers are even more notable when approached from a text-linguistic perspective.

4.2 *Grammar from the perspective of text-linguistic variation*

Multi-dimensional studies (e.g., Biber 1988, 1995, 2006) have used a text-linguistic approach to describe the grammatical characteristics of different texts and registers. One of the most important findings from these studies is the extent to which spoken registers rely on clausal grammar, while informational/academic written registers rely on phrasal grammar. For example, Dimension 1 in both the 1988 study of general spoken and written registers, as well as the 2006 study of university spoken and written registers, shows this same opposition: between verbs and finite dependent clauses (e.g., conditional clauses and WH-clauses) co-occurring frequently with pronouns and adverbials in spoken registers, versus nouns, attributive adjectives, and prepositional phrases co-occurring frequently in written academic registers.

Several other studies have documented this fundamental difference between spoken and informational-written discourse in more detail (Biber and Clark 2002; Biber 2009a). For example, Biber and Gray (2010) use a text-linguistic design to analyze the rates of occurrence for clausal versus phrasal features associated with grammatical complexity (i.e., dependent clause types versus phrasal modifiers in noun phrases). That study contrasts the patterns of use in conversation versus academic research writing, based on analysis of a 4-million-word corpus of American English (AmE) conversation (723 texts) and a 3-million-word corpus of academic research articles (429 texts). Two structural factors turn out to be important when accounting for the differences between these two registers:

1. structural type:
   a. clauses, especially finite dependent clauses, are preferred in speech
   b. (non-verbal) phrases are preferred in academic writing
2. syntactic function:
   a. clausal constituents (adverbials and complement clauses) are preferred in speech
   b. noun phrase constituents (noun modifiers and noun complements) are preferred in academic writing

Tables 2 and 3 provide selected details of these general trends, comparing the mean scores for many of the most important complexity features.

In sum, from a text-linguistic perspective, the grammar of conversation is dramatically different from the grammar of informational writing (cf. Biber, Gray, and Poonpon 2011). This does not represent an absolute difference between speech and writing, because written registers like email can employ the grammatical discourse styles typical of speech (e.g., Biber and Conrad 2009: Ch. 7). However, informational writing is fundamentally different from conversation (and spoken registers generally) in its heavy reliance on phrasal

Table 2. *Grammatical complexity features typical of conversation*

Linguistic feature	Conversation mean score	Academic WR mean score	*F* value	significance	r^2
Finite adverbial clauses					
total adverbial clauses	7.1	3.6	603.2	<.0001	.35
Verb complement clauses					
verb + *that*-clause	10.8	2.6	2196.7	<.0001	.66
verb + *WH*-clause	2.7	0.2	1413.9	<.0001	.55
verb + *ing*-clause	1.3	0.2	842.5	<.0001	.42
verb + *to*-clause	4.7	3.4	166.6	<.0001	.13

Table 3. *Grammatical complexity features typical of academic writing*

Linguistic feature	Conversation mean score	Academic WR mean score	*F* value	significance	r^2
Noun modifiers – clauses					
WH relative clauses	0.9	3.7	858.1	<.0001	.43
nonfinite relative clauses	0.7	4.2	2257.3	<.0001	.66
Noun complement clauses					
noun + *that*-clause	0.1	0.6	474.1	<.0001	.29
noun + *to*-clause	0.9	2.8	856.8	<.0001	.43
Noun modifiers – phrasal					
Attributive adjectives	16.5	57.1	5787.8	<.0001	.84
Nouns as nominal pre-mods	19.0	57.4	1259.2	<.0001	.52
Total prepositional phrases as nominal post-mods	6.3	51.9	1380.1	<.0001	.94

rather than clausal modification. Thus, any description of variation in English that disregards this opposition will miss out on what is arguably the most important predictor of grammatical differences.

5. Lexis, grammar, and register

5.1 *Variationist and text-linguistic perspectives on the lexis-grammar interface*

The *Longman Grammar of Spoken and Written English* (LGSWE by Biber et al. 1999) provides extensive descriptive information on the interaction of lexis and grammar. Lists of the most common words occurring with many grammatical constructions are grouped according to their specific frequency band, including: phrasal verbs (pp. 409–412); prepositional verbs (pp. 416–421); perfect aspect verbs (pp. 463–464); passive voice verbs (pp. 477–480);

attributive adjectives (pp. 511–515); predicative adjectives (pp. 515–518); nouns controlling complement clauses (pp. 648–653); verbs controlling a complement clauses (pp. 667–670, pp. 688–693, pp. 710–714, pp. 741–748); adjectives controlling complement clauses (pp. 672–673, pp. 718–721, p. 749); stance adverbials (pp. 867–871); and linking adverbials (pp. 886–889). Other kinds of quantitative information are also provided, such as the extent to which individual verbs and particular particles/prepositions co-occur in phrasal verbs (pp. 412–413) or in prepositional verbs (pp. 422–423). Similarly, grammatical preferences are identified quantitatively for particular words, such as lists of verbs that usually occur in present tense or usually in past tense (p. 459); verbs that usually occur in progressive aspect or almost never occur in progressive aspect (pp. 471–472); and verbs that usually occur in passive voice or almost never occur in passive voice (pp. 477–482).

Both variationist and text-linguistic perspectives on the lexis-grammar interface are explored in the LGSWE:

1. Linguistic variation: how strongly is a given word associated with a grammatical construction?
2. Text-linguistic variation: how often is a lexical-grammatical combination encountered in natural texts and different registers?

As described above, the first perspective is described in terms of probabilistic preference, considering the proportional extent to which a word is used with a particular grammatical construction. For example, although the prepositional verb *base on* is not especially frequent in academic writing, it has an extremely strong association with passive voice, with more than 90% of all occurrences occurring as passives (LGSWE p. 479). In contrast, the second perspective can be described by computing rates of occurrence. For example, the verbs *be given, be found,* and *be seen* are the most frequently occurring passive verbs that a reader will encounter in academic writing (LGSWE p. 478).

Lexico-grammatical associations have been investigated in more detail through 'collostructional analysis', an advanced statistical approach that measures the strength of association between words and grammatical constructions (Stefanowitsch and Gries 2003). The methodological goal of collostructional analysis is to develop a statistical approach that does not assume a normal distribution or homogeneity of variance, and is robust for both rare and high-frequency collocations. Collostructional analysis achieves this goal by applying the Fisher exact test, which integrates both proportional preference and frequency rates of occurrence to produce a single measure that identifies the words that are most strongly associated with a target grammatical construction ('collexemes'). This type of analysis has been applied to numerous case studies, including an identification of the verbs that are most strongly attracted to past tense; the verbs that are most strongly attracted to progressive aspect;

the verbs that are most strongly attracted to active versus passive voice and the verbs that are most strongly attracted to the ditransitive versus *to*-dative (cf. Stefanowitsch and Gries 2003; Gries and Stefanowitsch 2004).

Collostructional analysis – similar to previous statistical measures of collocational association – assumes the desirability of a single statistical measure of linguistic importance (lexico-grammatical association). However, there are theoretical and practical reasons why it might be preferable to distinguish between the variationist and text-linguistic perspectives: the words that are proportionally associated with a grammatical construction (even if they rarely occur), versus the words that most frequently occur with a grammatical construction (even if those words also frequently have other grammatical associations).

The two types of analysis produce results that are to a large extent complementary. For example, from the text-linguistic perspective, passive voice constructions in academic writing are most frequently found with verbs like *be made, be given, be used, be seen*, even though those verbs are also commonly used in the active voice (compare LGSWE pp. 367–369 and p. 375 with p. 478). In contrast, the variationist perspective identifies a completely different set of verbs have the strongest proportional association with passive voice in academic writing, occurring as passives over 90% of the time, even though none of them is especially frequent: *be aligned (with), based on, coupled with, deemed, effected, situated, subjected (to)*, etc. (LGSWE p. 479). Thus, a reader will most often encounter passive verbs like *be made, be given, be used*, and she/he will develop the association that when passive voice is employed, it often incorporates those verbs. In contrast, a reader will less often encounter passive verbs like *be aligned (with)* and *based on*, but because those verbs are almost never encountered in the active voice, an association in the opposite direction is developed: when those verbs are used, they are almost always in the passive voice. The two perspectives are distinct but both are important.

5.2 *Register influences on lexis and grammar*

Many previous studies of lexico-grammatical associations have disregarded register differences. For example, probably the most developed exploration of the lexis-grammar interface is the series of reference books and academic studies carried out under the rubric of 'pattern grammar'. Two major reference books have emerged from this framework (cf. Francis et al. 1996, 1998; Hunston and Francis 2000). These studies investigate the phraseology of individual words, showing 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. However, there is no indication of register differences in these studies.

The omission of register analyses in the pattern grammar books can be justified by the magnitude of that project, compiling lists of 'patterns' for thousands of verbs, nouns, and adjectives. Given finite resources, the authors chose to increase their coverage of words, rather than investigating the possibility of different patterns occurring in different registers. However, this omission is also typical of most more restricted lexico-grammatical studies.

One exception to this generalization is Stefanowitsch and Gries (2008), who investigate the influence of spoken versus written mode on the associations between lexical items and 'constructions'. As described above, this study uses 'collostructional analysis', a statistical procedure that combines proportional preference and frequency rates of occurrence to produce a single measure of association between words and grammatical constructions.

For example, one case study in that paper investigates the verbs that are most strongly associated with passive voice in speech and writing. Three different statistical comparisons are used:

1) for each verb, contrasting the association with active voice versus passive voice, carried out separately for the spoken and written modes;
2) for each verb, contrasting the association with the spoken mode versus written mode, carried out separately for active voice and then passive voice.
3) for each verb, checking for cross-over effects: when a verb is associated with passive voice in the spoken mode but active voice in the written mode, or vice versa.

The first type of analysis identifies the set of verbs associated with passive (versus active) voice in speech, and then separately identifies the verbs associated with passive (versus active) voice in writing. (Stefanowitsch and Gries 2008: 140) Several verbs have associations with passive voice in both modes, such as: *BE + concerned, based, published, associated, confined, designed*. However, other verbs are associated with passive (versus active) voice only in one of the two modes.

In several cases, the results of this type of analysis are surprising, identifying verbs that would not normally be associated with the target register. In fact, many of these findings are difficult to interpret in terms of spoken versus written discourse, calling into question the value of this statistical measure for this application (cf. below). For example, the verbs most strongly associated with passive voice in speech (but not associated with passive voice in writing) include 'literate' verbs like: *BE + involved, used, engaged, enclosed, aimed, distributed, compared, entitled*. In contrast, the verbs most strongly associated with passive voice in writing (but not associated with passive voice in speech) include 'colloquial' verbs like *BE + thought, done, made*.

The second type of analysis identifies the set of verbs that are more strongly associated with passive voice in speech than in writing, and vice versa.

(Stefanowitsch and Gries 2008: 142). Here again, we find several seemingly 'literate' verbs in the 'spoken-passive' list, such as *BE + concerned, involved, cross-examined, readmitted, adduced, extruded*, while several 'colloquial' verbs are in the 'written-passive' list, such as *BE + had, known, got|gotten, thought, wanted*.

The third type of analysis identified only two verbs with cross-over effects: *find* and *work*, which are both attracted to spoken/active versus written/passive. However, these verbs are claimed to involve different meaning senses in the two channels, and thus to not represent a genuine register difference. Thus, the general conclusion that Stefanowitsch and Gries draw is that register differences are not important for this type of investigation: "the results of channel-sensitive collostructional analysis are essentially identical to those yielded by a 'channel-ignorant' analysis as far as constructional meaning in the narrow sense is concerned: we found no interaction between channel and semantics at all." (2008: 143)

Before addressing the claimed lack of a register effect, it is useful to discuss three other important issues that arise in the Stefanowitsch and Gries study. The first is to emphasize yet again the need to distinguish between the variationist and text-linguistic perspectives, and specifically to avoid claims about 'frequency' in variationist studies that investigate proportional preference. As noted in Section 2 above, this confusion arises even in some of the most carefully designed corpus-based studies. Stefanowitsch and Gries (2008) sometimes seem to fall into this same trap; for example, "the passive construction occurs relatively frequently with formal vocabulary in both channels." (p. 143) The normal interpretation of this statement is that a speaker would frequently produce passive constructions with formal vocabulary in speech: a text-linguistic perspective. However, Stefanowitsch and Gries (2008) did not actually consider this perspective and provide no findings to support such a conclusion. In fact, passives are not frequent at all in conversation. For example, the LGSWE (Figure 6.7; p. 476) shows that passive verbs occur only c. 2,000 times per million words in conversation, contrasted with c. 18,000 times per million words in academic writing. Further, the verbs that most frequently occur with passive voice in conversation are not formal vocabulary; rather, they are mostly everyday, colloquial verbs like *BE + made, done, called, put, told, born, paid* (cf. LGSWE pp. 478–479).

In contrast, the Stefanowitsch and Gries (2008) study apparently reflects the fact that when a 'formal' verb (like *engaged, enclosed, adduced,* or *extruded*) is used in conversation, it is usually used in the passive voice. But this does not at all mean that such combinations occur frequently in conversation.

The second general issue here is that surprising findings require interpretation and explanation; it is not sufficient to simply report surprising statistical results with no discussion. This is especially the case when we are relying on a complex

statistical measure, with results that are not necessarily well-understood. Such an analysis raises two possibilities: 1) that there are completely unanticipated linguistic patterns identified by the analysis, requiring a radical change in our understanding of spoken and written discourse; or 2) that the statistical approach is not measuring what we think it is, and thus the approach itself requires further analysis and interpretation.

For example, Table 6 in the Stefanowitsch and Gries (2008: 142) study lists the verbs that are most strongly associated with passive voice in speech contrasted with writing. The expectation of the reader is that these verbs are somehow especially typical of speech. The verbs identified by this method, though, are extremely surprising, including *be readmitted, be adduced, be extruded* – and no discussion is provided to explain how these passive verbs are typical of speech in contrast to writing. Similarly, no discussion is offered to explain what it means for verbs like *BE had, BE got|gotten, BE wanted* to be among the verbs most strongly associated with passive voice in writing (in contrast to speech).

Finally, these findings point to the need for triangulation of methodological approaches, and the risks of relying exclusively on any single measure. For example, multivariate statistics, simple descriptive statistics, and consideration of linguistic examples in textual contexts should all be considered and reconciled. Similarly, considering quantitative findings from both variationist and text-linguistic perspectives can provide a more complete description than any single perspective. Most importantly, analyses based on a single methodological approach should be presented with explicit discussion of the limitations of that approach.

For example, descriptive statistics on the use of passive verbs can be used to illustrate the complementary kinds of information found from the variationist versus text-linguistic analytical approaches. At the same time, these descriptive statistics identify some completely different patterns from those identified in the collostructional analysis. For example, an analysis of text-linguistic rate-of-occurrence identifies five verbs that are especially frequent with the passive voice in academic writing: *BE + found, given, made, seen, used*. (LGSWE p. 478). However, three of those verbs are not identified at all in the first collostructional analysis (verbs associated with passive voice versus active voice in academic writing): *BE + given, found, seen*. Similarly, *BE + given, found* are not included on the list of passive verbs associated with writing versus speech (even though they occur c. 100 times more frequently as passive verbs in academic writing than in speech; cf., LGSWE p. 478).

Other verbs that did not make it to the lists in the Stefanowitsch and Gries (2008) study have strong proportional use with passive voice in academic writing. For example, the following verbs were not identified in any of the collostructional analyses, but descriptive statistics show that they all occur as

Table 4. *Selected verbs showing different patterns for frequency rates of occurrence (per 1 million words) and proportional use of passive voice, comparing conversation (AmE) and academic writing* (based on analysis of c. 5-million word corpora for each register, taken from the LSWE Corpus).

		Conversation		Academic Writing	
		Rate of Occurrence	Proportional Use of Passive Voice	Rate of Occurrence	Proportional Use of Passive Voice
Total for all main verbs (finite and non-finite)					
	Passive	14,000	2%	110,000	25%
	non-passive	645,000		429,000	
Passive associated with writing but not speech					
high frequency verbs; strong proportional preference					
Find	passive	26	0.8%	2,900	46%
	active	3,030		3,400	
high frequency; moderate proportional preference					
Show	passive	10	0.5%	2,310	28%
	active	2,090		6,000	
Make	passive	180	1.8%	2,500	30%
	active	10,200		5,800	
moderate frequency; strong proportional preference					
Call	passive	337	5.5%	1,040	40%
	active	5,800		1,553	
Passive associated with speech but not writing					
low frequency verbs; very strong proportional preference					
Allow	passive	170	57%	190	7.5%
	active	130		2,350	
Stick	all passive	33	63%	2	21%
	(get passive	10)			
	active	19		8	

passives in academic writing over 90% of the time: *BE + aligned, coupled with, deemed, effected, flattened, inclined, obliged, positioned, situated, stained, subjected to*.

When descriptive statistics are considered, and both proportional/variationist as well as text-linguistic analyses are employed, it is easy to isolate major differences between spoken and written discourse. For example, the quantitative distributions shown in Table 4 for verbs in the passive versus active voice illustrate many of the possible patterns of use, such as very frequent but weak proportional preference; infrequent but strong proportional preference; etc. In addition, these case studies illustrate strong differences between conversation and academic writing, from both analytical perspectives.

The descriptive findings presented in Table 4 show that register is a fundamentally important factor for the description of lexico-grammatical associations,

leading to different conclusions from the collostructional analysis of passive constructions. First of all, the descriptive statistics show that register has a major influence, while the collostructional analysis concluded that register had essentially no influence. Beyond that, the lists of specific verbs associated with the passive in speech or writing are strikingly different in the two analyses. Finally, the descriptive statistics show that the variationist and text-linguistic perspectives are clearly independent, indicating that they should be analyzed separately.

It is not completely clear how to reconcile the quantitative results from the Stefanowitsch and Gries (2008) collostructional analysis with these descriptive statistics. However, this is an essential task: to be useful, statistical analyses must be interpretable in linguistic terms. In the case study presented above, it is difficult to interpret the collostructional results from the perspective of a normal conversational interlocutor or a typical reader of written texts. For example, many of the passive verbs with high distinctiveness scores for speech are clearly not typical of the verbs that a speaker would normally encounter in conversation (e.g., *be readmitted, be adduced, be extruded*). And conversely, many of the most frequent passive verbs that a listener/reader would encounter are disregarded by the collostructional analysis (e.g., the verbs *be given, be found, be seen* in academic writing). Statistical analyses of corpora often uncover patterns of language use that we had not noticed before. But we should subsequently be able to confirm the results of the statistical analysis in natural texts: there should not be a disconnect between the language that we observe in texts and the results of our statistical analysis.

Beyond that, there is reason to question the desirability of a single measure of lexical association. In particular, the most frequent patterns found in texts are often not the same as the proportionally preferred associations. It is convenient to have a single score that measures association strength. However, it is not clear that that approach generally provides the most useful description of actual language use.

In sum, the main point of this section has been to argue for the importance of register differences in describing lexico-grammatical patterns of use. The study by Stefanowitsch and Gries (2008) has been discussed in some detail because it is one of the few previous studies to investigate register influences on lexico-grammatical associations. The omnibus measure of collostructional strength employed in that study suggests that register is not a strong factor influencing lexico-grammatical variation, or at least that register "does not interact substantially with constructional semantics." (p. 129) In contrast, I have argued here that a separate consideration of distributional patterns associated with the variationist versus text-linguistic perspectives both show consistent and important differences in lexico-grammatical patterns across registers.

6. Conclusion

The present paper argues for the importance of register differences at all linguistic levels. However, as background, the paper first distinguishes between two major approaches to the study of linguistic variation and use: variationist versus text-linguistic. The variationist approach has been widely used since the 1960's for the quantitative study of the 'sociolinguistic variable' (see, e.g., Labov 1972; Lavandera 1978). More recently, functional linguists from several different sub-disciplines have employed similar approaches to study grammatical variation, exploring the contextual factors that influence the choice among related grammatical variants. Variationist studies differ in their linguistic focus (phonological versus grammatical features) and in the statistical techniques that they employ (e.g., Varbul, logistic regression). However, all variationist studies share certain characteristics: 1) the research goal is to describe a linguistic feature, rather than the characteristics of texts; 2) each occurrence of the target linguistic feature constitutes an observation in the research design; and 3) the quantitative findings represent proportional preference for one linguistic variant in comparison to other variants.

The text-linguistic approach to linguistic variation differs in all three respects: 1) the research goal is to describe the characteristics of texts, rather than the characteristics of a linguistic feature; 2) each text constitutes an observation; and 3) the quantitative findings represent the rates of occurrence of linguistic features in texts rather than the proportional preference for a linguistic variant in comparison to other variants. Thus, the text-linguistic approach describes language use from the perspective of a conversational participant or a normal reader of a text: what features will they encounter most commonly in spoken interactions or written texts?

The description of passive voice verbs (cf. Section 5.2) illustrates the practical consequences of this distinction. From a variationist perspective, analyzing proportional preference, verbs like *BE + concerned, involved, cross-examined, readmitted, adduced, extruded* are especially associated with passive voice in speech. That is, when one of these verbs is used in spoken discourse (or at least in the particular corpus of speech analyzed for the S&G study), it is likely to occur as a passive rather than active voice verb. (In this case, these findings are in contrast to written discourse: these particular verbs are proportionally more likely to occur as passives in spoken discourse than in written discourse.)

The text-linguistic approach provides a dramatically different perspective, because most of these verbs are simply not common at all in spoken texts. Thus, a conversational participant will rarely encounter these verbs in text, whether they are in the active or passive voice. However, there are other verbs that do frequently occur with passive voice in conversation, like *BE + made, done, called, put, told, paid*. Proportionally, these verbs occur most of the time as

active voice verbs, so they are not associated with the passive from a variationist perspective. But a conversational participant will encounter these passive forms much more frequently than proportionally-preferred verbs like *BE adduced* or *BE extruded*.

Similar contrasts between the types of information provided by the two perspectives can be given for most linguistic features. The point here is not to argue that one or the other perspective is correct. However, studies often fail to distinguish between the two, slipping into descriptions that suggest a text-linguistic perspective when the data are strictly proportional or variationist. The secondary goal of the present paper has thus been to emphasize the difference between these two perspectives, and the need to explicitly characterize findings as relating to one or the other.

The primary goal, though, has been to argue for the importance of register differences – in both variationist and text-linguistic studies. That is, the patterns of linguistic variation and use are dramatically different across registers. The descriptions here have focused mostly on the spoken/written contrast (especially face-to-face conversation versus academic writing), but systematic differences exist across the full range of registers (see, e.g., Biber and Conrad 2009). These register differences exist across all linguistic levels, including lexical patterns, grammatical patterns, and lexico-grammatical associations.

Traditionally, most general-purpose corpora were designed to include multiple registers, and thus many descriptive studies have adopted a text-linguistic approach and include some information on register differences.[6] In recent years, some variationist studies have also begun to include analysis of register differences interacting with other factors of the linguistic context (see, e.g., Riordan 2007; Szmrecsanyi and Hinrichs 2008). However, it is still the norm in most studies of collocation and lexico-grammatical associations to disregard the possible influence of register differences. The main point of the present paper is that we should instead treat this possible influence as a likelihood: that the patterns of linguistic variation and use are usually strikingly different in spoken versus written registers. Thus, the practice advocated here is to begin a research study with the hypothesis that such register differences exist, and to include analysis of those differences unless they are empirically shown to be unimportant.

Bionotes

Douglas Biber is Regents' Professor of English (Applied Linguistics) at Northern Arizona University. His research efforts have focused on corpus linguistics, English grammar, and register variation (in English and cross-linguistic; synchronic and diachronic). He has written 170 research articles

and 13 books and monographs, including academic books published with Cambridge University Press (1988, 1995, 1998, 2009), John Benjamins (2006, 2007), the co-authored Longman Grammar of Spoken and Written English (1999), and three grammar textbooks published by Longman. Email: Douglas.Biber@nau.edu

Notes

1. In a few cases, corpus linguists have actively argued against the need for quantitative analysis. One of the best known linguists to take this position is Sinclair, stating 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: 343–344).
2. It is also possible to include quantitative variables in a Type A study. For example, each occurrence of a relative clause could be coded for the number of words separating the relative pronoun from the gap position. However, studies of linguistic variation generally do not include such variables.
3. If the three sub-corpora had been exactly the same size, and if the analysis had been based on a complete sample of all *that*-clauses, then conclusions of this type would be appropriate. However, most Type A studies do not meet these two requirements.
4. Type C studies are also designed for text-linguistic research questions, but they do not permit the use of inferential statistics. The results reported in the *LGSWE* are actually based on Type C designs rather than Type B designs.
5. The operational definition of 'collocate' can also be approached from both variationist versus text-linguistic perspectives. The variationist perspective uses statistics like mutual information and log-likelihood, which are based on the proportion of both words that co-occur. The text-linguistic perspective uses simple rate of occurrence, measuring how often the combination of words is found in texts.
6. Most of these studies use a 'Type C' design, treating each sub-corpus as an observation rather than analyzing each text as an observation. While this still results in a text-linguistic perspective, it does not permit analysis of dispersion or the extent to which register differences hold for individual texts.

References

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, Douglas. 2006. *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, Douglas. 2009a. Are there linguistic consequences of literacy? Comparing the potentials of language use in speech and writing. In David R. Olson & Nancy Torrance (eds.), *Cambridge Handbook of Literacy*, 75–91. Cambridge: Cambridge University Press.

Biber, Douglas. 2009b. A corpus-driven approach to formulaic language: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14. 381–417.

Biber, Douglas & Victoria Clark. 2002. Historical shifts in modification patterns with complex noun phrase structures: How long can you go without a verb? In Teresa Fanego, Maria Jose Lopez-Couso & Javier Perez-Guerra (eds.), *English historical syntax and morphology*, 43–66. Amsterdam: John Benjamins.

Biber, Douglas & Susan Conrad. 2009. *Register, genre, and style*. Cambridge University Press.

Biber, Douglas, Susan Conrad & Viviana Cortes. 2004. If you look at . . . : Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25. 371–405.

Biber, Douglas, Susan Conrad & Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, Douglas & Bethany Gray. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes* 9. 2–20.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *The Longman grammar of spoken and written English*. London: Longman.

Biber, Douglas & James K. Jones. 2009. Quantitative methods in corpus linguistics. In Anke Lüdeling and Merja Kytö (eds.), *Corpus linguistics: An international handbook*, 1286–1304. Berlin: Walter de Gruyter.

Biber, Douglas, Bethany Gray, and Kornwipa Poonpon. 2011. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly* 45. 5–35.

Carter, Ronald & Michael McCarthy. 2006. *Cambridge grammar of English*. Cambridge: CUP.

Collins COBUILD English Grammar. 1990. London: Collins.

Conrad, Susan & Douglas Biber. 2009. *Real grammar: A corpus-based approach to English*. Pearson Longman.

Francis, Gill, Susan Hunston & Elizabeth Manning (eds.). 1996. *Collins COBUILD grammar patterns 1: Verbs*. London: HarperCollins.

Francis, Gill, Susan Hunston & Elizabeth Manning (eds.). 1998. *Collins COBUILD grammar patterns 2: Nouns and adjectives*. London: HarperCollins.

Gries, Stefan Th. 2006. Exploring variability within and between corpora: some methodological considerations. *Corpora* 1. 109–151.

Gries, Stefan Th. & Anatol Stefanowitsch. 2004. Extending collostructional analysis: a corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9. 97–129.

Hinrichs, Lars & Benedikt Szmrecsanyi. 2007. Recent changes in the function and frequency of standard English genitive constructions: a multivariate analysis of tagged corpora. *English Language and Linguistics* 11. 335–378.

Hunston, Susan. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunston, Susan & Gill Francis. 2000. *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Kennedy, Greame. 1991. *Between* and *through*: The company they keep and the functions they serve. In Karin Aijmer & Bengt Altenberg (eds.), *English Corpus Linguistics*, 95–110. London: Longman.

Kennedy, Graeme. 1998. *An introduction to corpus linguistics*. London: Longman.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Lavandera, Beatriz R. 1978. Where Does the Sociolinguistic Variable Stop? *Language in Society* 7. 171–82.

Longman Dictionary of Contemporary English. 2009. London: Longman.

Partington, Alan. 1998. *Patterns and meanings*. Amsterdam: John Benjamins.
Quirk, Randolf, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1972. *A grammar of contemporary English*. London: Longman.
Quirk, Randolf, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.
Riordan, Brian. 2007. There's two ways to say it: Modeling nonprestige there's. *Corpus Linguistics and Linguistic Theory* 3. 233–279.
Römer, Ute. 2005. *Progressives, patterns, pedagogy: A corpus-driven approach to English progressive forms, functions, contexts and didactics*. Amsterdam/Philadelphia: John Benjamins.
Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8. 209–243.
Stefanowitsch, Anatol & Stefan Th. Gries. 2008. Channel and constructional meaning: A collostructional case study. In Gitte Kristiansen & René Dirven (eds.), *Cognitive Sociolinguistics*, 129–152. Berlin & New York: Mouton de Gruyter.
Szmrecsanyi, Benedikt. 2005. Language users as creatures of habit: a corpus-linguistic analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1. 113–150.
Szmrecsanyi, Benedikt & Lars Hinrichs. 2008. Probabilistic determinants of genitive variation in spoken and written English: a multivariate comparison across time, space, and genres. In Nevalainen, Terttu, Irma Taavitsainen, Päivi Pahta & Minna Korhonen (eds.), *The dynamics of linguistic variation: Corpus evidence on English past and present*, 291–309. Amsterdam: Benjamins.