Corpus-Based and Corpus-driven Analyses of Language Variation and Use

Douglas Biber

The Oxford Handbook of Linguistic Analysis
Edited by Bernd Heine and Heiko Narrog

Print Publication Date: Dec 2009
Subject: Linguistics, Sociolinguistics, Morphology and Syntax
Online Publication Date: Sep 2012
DOI: 10.1093/oxfordhb/9780199544004.013.0008

Abstract and Keywords

Corpus linguistics is a research approach that has developed over the past few decades to support empirical investigations of language variation and use, resulting in research findings which have much greater generalizability and validity than would otherwise be feasible. Corpus studies have used two major research approaches: 'corpus-based' and 'corpus-driven'. Corpus-based research assumes the validity of linguistic forms and structures derived from linguistic theory. The primary goal of research is to analyse the systematic patterns of variation and use for those pre-defined linguistic features. Corpus-driven research is more inductive, so that the linguistic constructs themselves emerge from analysis of a corpus. This chapter illustrates the kinds of analyses and perspectives on language use possible from both corpus-based and corpus-driven approaches.

Keywords: corpus linguistics, language variation, language use, linguistic theory

8.1 Introduction

C□ □ □linguistics is a research approach that has developed over the past several decades to support empirical investigations of language variation and use, resulting in research findings that have much greater generalizability and validity than would otherwise be feasible. Corpus linguistics is not in itself a model of language. In fact, at one level it can be regarded as primarily a methodological approach:

- it is empirical, analyzing the actual patterns of use in natural texts;

- it utilizes a large and principled collection of natural texts, known as a "corpus", as the basis for analysis;

(p. 160) • it makes extensive use of computers for analysis, using both automatic and interactive techniques;

- it depends on both quantitative and qualitative analytical techniques (Biber et al. 1998: 4).

At the same time, corpus linguistics is much more than a methodological approach: these methodological innovations have enabled researchers to ask fundamentally different kinds of research questions, sometimes resulting in radically different perspectives on language variation and use from those taken in previous research. Corpus linguistic research offers strong support for the view that language variation is systematic and can be described using empirical, quantitative methods. Variation often involves complex patterns consisting of the interaction among several different linguistic parameters, but, in the end, it is systematic. Beyond this, the major contribution of corpus linguistics is to document the existence of linguistic constructs that are not recognized by current linguistic theories. Research of this type—referred to as a "corpus-driven" approach—identifies strong tendencies for words and grammatical constructions to pattern together in particular ways, while other theoretically possible combinations rarely occur. Corpus-driven research has shown that these tendencies are much stronger and more pervasive than previously suspected and that they usually have semantic or functional associations (see section 8.3 below).

In some ways, corpus research can be seen as a logical extension of quantitative research in sociolinguistics begun in the 1960s (e.g., Labov 1966), which rejected "free variation" as an adequate account of linguistic choice and argued instead for the existence of linguistic variable rules (see Chambers and Trudgill 1980: 59–61; 146–9). However, research in corpus linguistics differs from quantitative sociolinguistic research in at least two major ways:

(1) Quantitative sociolinguistics has focused on a relatively small range of varieties: usually the social dialects that exist within a single city, with secondary attention given to the set of "styles" that occur during a sociolinguistic interview. In contrast, corpus research has investigated the patterns of variation among a much wider range of varieties, including spoken and written registers as well as dialects.

Corpus-based dialect studies have investigated national varieties, regional dialects within a country, and social dialects. However, the biggest difference from quantitative sociolinguistics here has to do with the investigation of situationally-defined varieties: "registers". Quantitative sociolinguistics has restricted itself to the investigation of only

spoken varieties, and considered only a few “styles”, which speakers produce during the course of a sociolinguistic interview (e.g., telling a story vs. reading a word list). In contrast, corpus-based research investigates the patterns of variation among the full set of spoken and written registers in a language. In speech, these include casual face-to-face conversation, service (p. 161) encounters, lectures, sermons, political debates, etc.; and, in writing, these include email messages, text-messaging, newspaper editorials, academic research articles, etc.

(2) Quantitative sociolinguistics has focused on analysis of “linguistic variables”, defined such that the variants must have identical referential meaning. Related to this restriction, quantitative sociolinguistic research has focused exclusively on nonfunctional variation. For these reasons, most quantitative sociolinguistic research has focused on phonological variables, such as [t] vs. [θ]. Sociolinguistic variation is described as indexing different social varieties, but there is no possibility of functional explanations for why a particular linguistic variant would be preferred in one variety over another.

In contrast, corpus research considers all aspects of language variation and choice, including the choice among roughly synonymous words (e.g., *big, large, great*), and the choice among related grammatical constructions (e.g., active vs. passive voice, dative movement, particle movement with phrasal verbs, extraposed vs. subject complement clauses). Corpus-based research goes even further, investigating distributional differences in the extent to which varieties rely on core grammatical features (e.g., the relative frequency of nouns, verbs, prepositional phrases, etc.). All of these aspects of linguistic variation are interpreted in functional terms, attempting to explain the linguistic patterns by reference to communicative and situational differences among the varieties. In fact, much corpus-based research is based on the premise that language variation is functional: that we choose to use particular linguistic features because those forms fit the communicative context of the text, whether in conversation, a political speech, a newspaper editorial, or an academic research article.

In both of these regards, corpus-based research is actually more similar to research in functional linguistics than research in quantitative sociolinguistics. By studying linguistic variation in naturally occurring discourse, functional linguists have been able to identify systematic differences in the use of linguistic variants. An early study of this type is Prince (1978), who compares the distribution and discourse functions of WH-clefts and *it*-clefts in spoken and written texts. Thompson and Schiffrin have carried out numerous studies in this research tradition: Thompson on detached participial clauses (1983), adverbial purpose clauses (1985), omission of the complementizer *that* (Thompson and Mulac 1991*a*; 1991*b*), relative clauses (Fox and Thompson 1990); and Schiffrin on verb tense (1981), causal sequences (1985*a*), and discourse markers (1985*b*). Other early

studies of this type include Ward (1990) on VP preposing, Collins (1995) on dative alternation, and Myhill (1995; 1997) on modal verbs.

More recently, researchers on discourse and grammar have begun to use the tools and techniques available from corpus linguistics, with its greater emphasis on the representativeness of the language sample, and its computational tools for (p. 162) investigating distributional patterns across registers and across discourse contexts in large text collections (see Biber et al. 1998; Kennedy 1998; Meyer 2002; and McEnery et al. 2006). There are a number of book-length treatments reporting corpus-based investigations of grammar and discourse: for example, Tottie (1991*a*) on negation, Collins (1991) on clefts, Mair (1990) on infinitival complement clauses, Meyer (1992) on apposition, Mindt 1995 on modal verbs, Hunston and Francis (2000) on pattern grammar, Aijmer (2002) on discourse particles, Rohdenburg and Mondorf (2003) on grammatical variation; Lindquist and Mair (2004) on grammaticalization, Mahlberg (2005) on general nouns, Römer (2005) on progressives.

A central concern for corpus-based studies is the representativeness of the corpus (see Biber 1993; Biber et al. 1998: 246–50; McEnery et al. 2006: 13–21, 125–30). Two considerations are crucial for corpus design: size and composition. First, corpora need to be large enough to accurately represent the distribution of linguistic features. Second, the texts in a corpus must be deliberately sampled to represent the registers in the target domain of use.

Corpus studies have used two major research approaches: “corpus-based” and “corpus-driven”. Corpus-based research assumes the validity of linguistic forms and structures derived from linguistic theory; the primary goal of research is to analyze the systematic patterns of variation and use for those predefined linguistic features. One of the major general findings from corpus-based research is that descriptions of grammatical variation and use are usually not valid for the language as a whole. Rather, characteristics of the textual environment interact with register differences, so that strong patterns in one register often represent weak patterns in other registers. As a result, most corpus-based studies of grammatical variation include consideration of register differences. The recent *Longman Grammar of Spoken and Written English* (Biber et al. 1999) is the most comprehensive reference work of this kind, applying corpus-based analyses to show how any grammatical feature can be described for its patterns of use across discourse contexts and across spoken and written registers.

In contrast, “corpus-driven” research is more inductive, so that the linguistic constructs themselves emerge from analysis of a corpus. The availability of very large, representative corpora, combined with computational tools for analysis, make it possible to approach linguistic variation from this radically different perspective. The corpus-

driven approach differs from the standard practice of linguistics in that it makes minimal a priori assumptions regarding the linguistic features that should be employed for the corpus analysis. In its most basic form, corpus-driven analysis assumes only the existence of words, while concepts like “phrase” and “clause” have no a priori status. Rather, co-occurrence patterns among words, discovered from the corpus analysis, are the basis for subsequent linguistic descriptions.

The following sections illustrate the kinds of analyses and perspectives on language use possible from both corpus-based and corpus-driven approaches. (p. 163) section 8.2 illustrates the corpus-based approach, which documents the systematic patterns of language use, often showing that intuitions about use are wrong. section 8.3 then illustrates the corpus-driven approach, showing how corpus research can uncover linguistic units that are not detectable using the standard methods of linguistic analysis.

8.2 Corpus-based research studies

As noted above, the corpus-based approach has some of the same basic goals as research in functional linguistics generally, to describe and explain linguistic patterns of variation and use. The goal is not to discover new linguistic features but rather to discover the systematic patterns of use that govern the linguistic features recognized by standard linguistic theory.

One major contribution of the corpus-based approach is that it establishes the centrality of register for descriptions of language use. That is, corpus-based research has shown that almost any linguistic feature or variant is distributed and used in dramatically different ways across different registers. Taken together, corpus-based studies challenge the utility of general linguistic descriptions of a language; rather, these studies have shown that any linguistic description that disregards register is incomplete or sometimes even misleading.

Considered within the larger context of quantitative social science research, the major strengths of the corpus-based approach are its high reliability and external validity. The use of computational tools ensures high reliability, since a computer program should make the same analytical decision every time it encounters the same linguistic phenomenon. More importantly, the corpus itself is deliberately constructed and evaluated for the extent to which it represents the target domain (e.g., a register or dialect). Thus, the linguistic patterns of use described in corpus-based analysis are generalizable, explicitly addressing issues of external validity.

However, judged by the normal interests of linguists, the greater contribution of the corpus-based approach is that it often produces surprising findings that run directly counter to our prior intuitions. That is, as linguists we often have strong intuitions about language use (in addition to intuitions about grammaticality), believing that we have a good sense of what is normal in discourse. While it is difficult to evaluate intuitions about grammaticality, intuitions about use are open to empirical investigation. Corpus-based research is ideally suited for this task, since one of the main research goals of this approach is to empirically identify the linguistic patterns that are extremely frequent or rare in discourse from a particular (p. 164) variety. And when such empirical investigations are conducted, they often reveal patterns that are directly counter to our prior expectations.

A simple case study of this type, taken from the *Longman Grammar of Spoken and Written English* (Biber et al. 1999: 460–3), concerns the distribution of verb aspect in English conversation. There are three aspects distinguished in English verb phrases:

Simple aspect: Do you like it?

Progressive aspect: I was running around the house like a maniac.

Perfect aspect: You haven't even gone yet.

The question to consider is which grammatical aspect is most common in face-to-face conversation?
It is much easier to illustrate the unreliability of intuitions in a spoken lecture because audience members can be forced to commit to an answer before seeing the corpus findings. For full effect, the reader here should concretely decide on an answer before reading further.

Hundreds of linguists have been polled on this question, and the overwhelming majority have selected progressive aspect as the most common verb aspect in English conversation. In fact, as Figure 8.1 shows, progressive aspect is more common in conversation than in other registers. The contrast with academic prose is especially noteworthy: progressive aspect is rare in academic prose but common in conversation.

However, as Figure 8.2 shows, it is not at all correct to conclude that progressive aspect is the most common choice in conversation. Rather, simple aspect is clearly the unmarked choice. In fact, simple aspect verb phrases are more than 20 times as common as progressives in conversation.

The following conversation illustrates this extreme reliance on simple aspect (underlined) in contrast to the much more specialized use of progressive aspect (in ***bold italics***):

J Well girls we better open the presents, I'm going to fall asleep.

K I know.

A Okay, right after he rolls out this last batch.

R Your face is really hot. Why are you leaving it, we're not leaving till Sunday are we?

J Which ever day you prefer, Saturday or Sunday.

R When are you leaving?

A Sunday morning.

R Oh, well we don't have to do it right away.

K Oh well let's just do it.

R I'd rather wait till I feel like it.

J But we're doing it. Figure 8.1. Distribution of progressive aspect verb phrases across registers

K Just do and be done with it. Smoke a joint <laugh>.

J Rita that'd help you sleep.

R No

J I don't think so.

A They used to make me sleep.

R No that would make my mind race, yeah, typical.

J Okay let's do the Christmas.

R If I drink

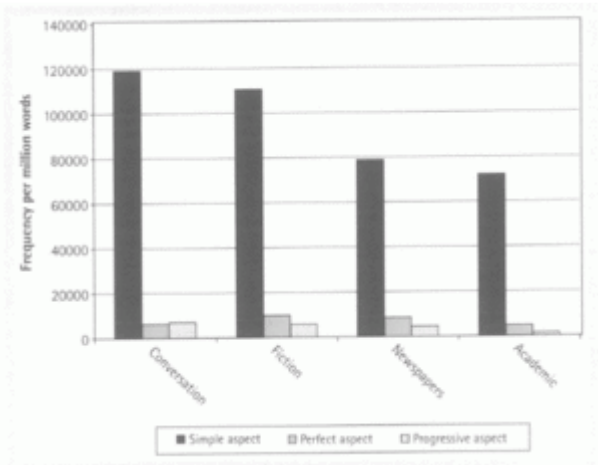
A Okay.

R If I smoke, anything, makes my mind race.

A These tins are the last ones.

J: It's just a little something Rita.

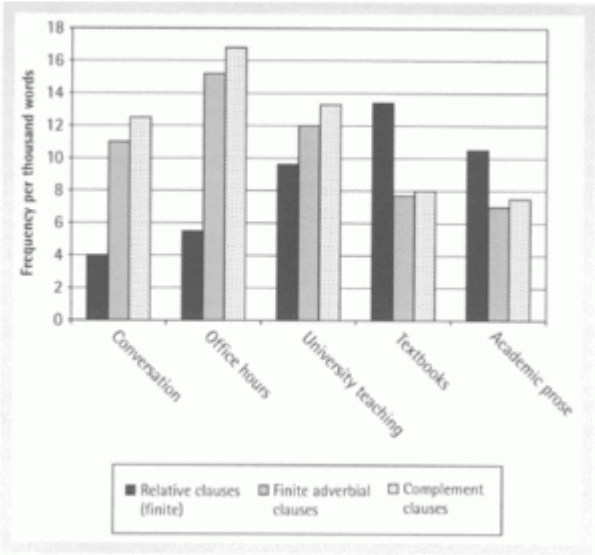
R: You go overboard. Now, don't you make us feel guilty.



Click to view larger

Figure 8.2. **Distribution of aspect types across registers**

As the conversational excerpt above shows, verbs of all types tend to occur with simple aspect rather than progressive aspect, including stative relational verbs (e.g., *be*), mental verbs (e.g., *know, prefer, feel, think*), verbs of facilitation or causation (p. 166) (e.g., *let, help, make*), and activity verbs (e.g., *do, open, fall, roll, wait, smoke, sleep, race, drink, go*). There are a few particular verbs that occur more often with progressive aspect than simple aspect, such as *bleeding, chasing, shopping, dancing, dripping, marching, raining, sweating, chatting, joking, moaning, looking forward to, studying, lurking* (see Biber et al. 1999: 471–5). However, the normal style of discourse in conversation relies on simple aspect verbs (usually present tense), with shifts into progressive aspect being used to mark specialized meanings.



Click to view larger

Figure 8.3. **Distribution of dependent clause types across registers**

A second case study—focusing on dependent clause types—illustrates how corpus-based research has established the centrality of register for descriptions of language use. Dependent clauses are often regarded as one of the best measures of grammatical complexity. In some approaches, all dependent clause types are grouped together as manifesting complexity, as with the use of t-unit length to measure language development. Further, there is a strong expectation that writing manifests a much greater use of dependent clauses than speech. So, for example, students are expected to develop increasing use of dependent clauses as they progress in their academic writing skills (see, for example, Wolfe-Quintero et al. 1998). (p. 167)

Corpus-based research has shown that these predictions are based on faulty intuitions about use. That is, different dependent clause types are used and distributed in dramatically different ways, and some dependent clause types are actually much more common in conversation than in academic writing. Thus, the practice of treating all types of dependent clause as a single unified construct has no basis in actual language use.

For example, Figure 8.3 compares the use of dependent clause types in five spoken and written registers: conversation, university office hours, university teaching, university textbooks, and academic prose. Relative clauses follow the expected pattern of being much more common in academic writing and textbooks than in conversation (and office hours). Class teaching is intermediate between conversation and academic writing in the use of relative clauses. However, the other two clause types—adverbial clauses and complement clauses—are much more common in conversation than in academic writing. Office hours are interesting here because they are even more sharply distinguished from writing, with extremely frequent use of adverbial clauses and complement clauses. Class

teaching is very similar to conversation in the frequent use of complement clauses and finite adverbial clauses.

(p. 168) Closer consideration of these patterns shows that they are interpretable in functional terms. For example, in conversation both adverbial and complement clauses occur with a highly restricted range of forms. Most adverbial clauses in conversation are finite, with especially high frequencies of *if*-clauses and *because*-clauses. Similarly, most complement clauses in conversation are finite (*that*-clauses and WH-clauses). In most cases, these complement clauses are controlled by a verb that expresses a "stance" relative to the proposition contained in the complement clause (e.g., *I thought that …, I don't know why …*).

In general, these distributional patterns conform to the general reliance on clausal rather than phrasal syntax in conversation (see Biber and Conrad to appear) and the communicative purposes of focusing on personal experience and activities rather than conveying more abstract information. These kinds of findings are typical of other corpus-based research, showing how the patterns of linguistic variation are systematically distributed in ways that have clear functional interpretations but are often not anticipated ahead of time.

8.3 Corpus-driven research studies

While corpus-based studies uncover surprising patterns of variation, corpus-driven analyses exploit 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 and Elena 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 and Elena 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. So, for example, Stubbs (1993: 16) cites the example of *eye* vs. *eyes*, taken from Sinclair (1991*b*). The plural form *eyes* often refers to the physical body part and is modified by an attributive adjective (e.g., *blue eyes*) or a possessive determiner (e.g., *your eyes*). In contrast, the singular form rarely refers to a specific body part but is commonly used in

fixed expressions, like *make eye contact, keep an eye on/out, catch your eye, in my mind's eye.* Thus, some corpus-driven research has challenged the utility of the notion of lemma, arguing instead that each word form tends to occur in distinctive grammatical contexts and tends to have distinct meanings and uses.

(p. 169) 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 corpus-based research by the nature of their central research goals:

- corpus-driven research: attempting to uncover new linguistic constructs through inductive analysis of corpora;
- corpus-based research: attempting to describe the systematic patterns of variation and use for linguistic features and constructs that have been previously identified by linguistic theory.

However, corpus-driven methodologies can differ from one study to the next in three key respects:

- the extent to which they are based on analysis of lemmas vs. each word form;
- the extent to which they are based on previously defined linguistic constructs (e.g., part-of-speech categories and syntactic structures) vs. simple sequences of words;
- the role of frequency evidence in the analysis.

The following sections survey some major corpus-driven studies, introducing the contributions that result from this research approach while also describing the key methodological differences within this general approach. section 8.3.1 illustrates one specific type of analysis undertaken from an extreme corpus-driven approach: the investigation of “lexical bundles”, which are the most common recurrent sequences of word forms in a register. It turns out that these word sequences have distinctive structural and functional correlates, even though they rarely correspond to complete linguistic structures recognized by current linguistic theories.

Next, section 8.3.2 surveys research done within the framework of “pattern grammar”. These studies adopt a more hybrid approach: they assume the existence of some grammatical classes (e.g., verb, noun) and basic syntactic structures, but they are corpus-driven in that they focus on the linguistic units that emerge from corpus analysis, with a primary focus on the inter-relation of words, grammar, and meaning. Frequency plays a relatively minor role in analyses done within this framework. In fact, as discussed in section 8.3.3, there is somewhat of a disconnect between theoretical discussions of the corpus-driven approach, where analyses are based on “recurrent patterns” and “frequency distributions” (Tognini-Bonelli 2001: 87), and the actual practice of scholars

working in pattern grammar, which has focused much more on form—meaning associations with relatively little accountability to quantitative evidence from the corpus.

Finally, section 8.3.4 introduces Multi-Dimensional analysis, which might also be considered a hybrid approach: it assumes the validity of predefined grammatical categories (e.g., nominalizations, past tense verbs) and syntactic features (e.g., WH relative clauses, conditional adverbial clauses), but it uses frequency-based corpus-driven methods to discover the underlying parameters of linguistic variation that best distinguish among spoken and written registers.

(p. 170) 8.3.1 Lexical bundles

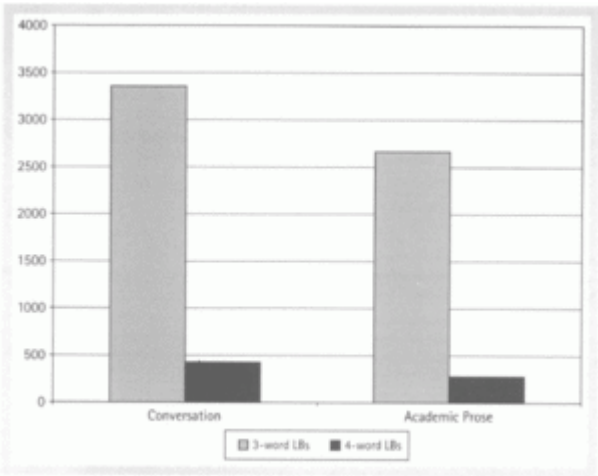
As noted above, the strictest form of corpus-driven analysis assumes only the existence of word forms. Some researchers interested in the study of formulaic language have adopted this approach, beginning with simple word forms and giving priority to frequency, to identify recurrent word sequences (e.g., Salem 1987; Altenberg and Eeg-Olofsson 1990; Altenberg 1998; Butler 1998; and Schmitt et al. 2004). Several of these studies have investigated recurrent word sequences under the rubric of "lexical bundles", comparing their characteristics in different spoken and written registers (e.g., Biber et al. 1999, Chapter 13; Biber and Conrad 1999; Biber et al. 2004; Cortes 2002; 2004; Partington and Morley 2004; Nesi and Basturkmen 2006; Biber and Barbieri 2007; Tracy-Ventura et al. 2007; and Biber et al. to appear).

Lexical bundles are defined as the multi-word sequences that recur most frequently and are distributed widely across different texts. Lexical bundles in English conversation are word sequences like *I don't know if* or *I just wanted to.* They are usually neither structurally complete nor idiomatic in meaning.

The initial analysis of lexical bundles in English (Biber et al. 1999, Chapter 13) compared the frequent word sequences in conversation and academic prose, based on analysis of *c.* 5-million-word sub-corpora from each register. Figure 8.4 shows the overall distribution of all 3-word and 4-word lexical bundles occurring more than 10 times per million words (distributed across at least five different texts). Not surprisingly, there are almost 10 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.

Lexical bundles are identified using a corpus-driven approach, based solely on distributional criteria (rate of occurrence of word sequences and their distribution across texts). As a result, lexical bundles are not necessarily complete structural units recognized by current linguistic theories. However, once they have been identified using

corpus-driven techniques, it is possible to carry out an interpretive analysis to determine if they have any systematic structural and functional characteristics.



Click to view larger

Figure 8.4. **Number of different lexical bundles in English (occurring more than 10 times per million words)**

This *post-hoc* analysis shows that lexical bundles differ from the formulaic expressions identified using traditional methods in three major respects. First, lexical bundles are by definition extremely common. 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 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 (p. 171) the beginning elements of a second structural unit. 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*).

In contrast, the formulaic expressions recognized by linguistic theory are usually complete structural units and idiomatic in meaning. However, corpus analysis shows that formulaic expressions with those characteristics are usually quite rare. For example, idioms such as *kick the bucket* and *a slap in the face* are rarely attested in natural conversation. (Idioms are occasionally used in fictional dialogue, but even there they are not common; see Biber et al. 1999: 1024–6).

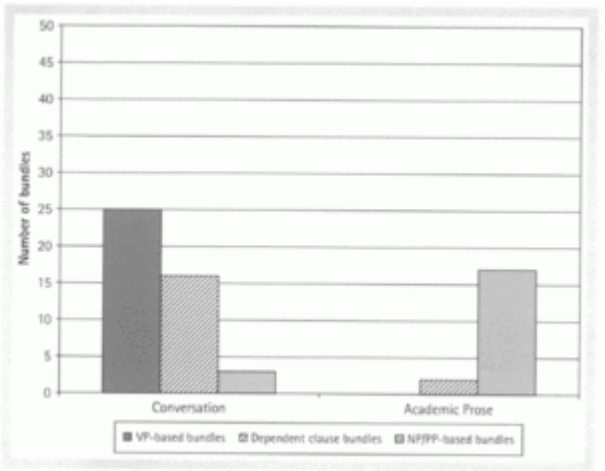
Although most lexical bundles are not complete structural units, they do usually have strong grammatical correlates. For example, bundles like *you want me to* are constructed

from verbs and clause components, while bundles like *in the case of* are constructed from noun phrase and prepositional phrase components. In English, two major structural types of lexical bundle can be distinguished: clausal and phrasal. Many clausal bundles simply incorporate verb phrase fragments, such as *it's going to be* and *what do you think.* Other clausal bundles are composed of dependent clause fragments rather than simple verb phrase fragments, such as *when we get* (p. 172) *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 8.5 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. (Figure 8.5 is based on a detailed analysis of the 4-word bundles that occur more than 40 times per million words.) In conversation, almost 90% of all common lexical bundles are declarative or interrogative clause segments. 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*). An additional 19% of the bundles consist of an extended verb phrase fragment (e.g., *have a look at*), while another 17% of the bundles are question fragments (e.g., *can I have a*). In contrast, the lexical bundles in academic prose are phrasal rather than clausal. Almost 70% of the common bundles in academic prose consist of a noun phrase with an embedded prepositional phrase fragment (e.g., *the nature of the*) or a sequence that bridges across two prepositional phrases (e.g., *as a result of*).

Although they are neither idiomatic nor structurally complete, lexical bundles are important building blocks in discourse. Lexical bundles often provide a kind of pragmatic "head" for larger phrases and clauses; the bundle functions as a discourse frame for the expression of new information in the following slot. That is, the lexical bundle usually 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,

> *I want you to write a very brief summary of his lecture.*
> *Hermeneutic efforts are provoked by the fact that the interweaving of system integration and social integration […] keeps societal processes transparent …*



Click to view larger

Figure 8.5. **Distribution of lexical bundles across structural types (4-word bundles occurring more than 40 times per million words)**

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

> ***Epistemic lexical bundles***:
> *I don't know what the voltage is here.*
> *I thought it was the other way around.*
> ***Attitudinal/modality bundles***:
> *I don't want to deliver bad news to her.*
> *All you have to do is work on it.*

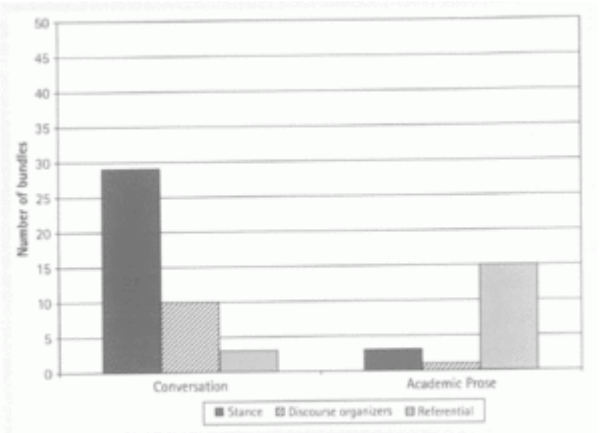
(p. 173)

Discourse-organizing bundles function to indicate the overall discourse structure: introducing topics, topic elaboration/clarification, confirmation checks, etc.:

> *What I want to do is quickly run through the exercise …*
> *Yes, you know there was more of a playful thing with it, you know what I mean?*

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

> *Students must define and constantly refine the nature of the problem.*
> *She's in that office down there, at the end of the hall.*



Click to view larger

Figure 8.6. **Distribution of lexical bundles across functional types (4-word bundles occurring more than 40 times per million words)**

Figure 8.6 shows that the typical discourse functions of lexical bundles are strikingly different in conversation vs. 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. It further turns out (p. 174) that there is a strong association between structural type and functional type for these lexical bundles: most stance bundles employ verbs or clause fragments, while most referential bundles are composed of noun phrase and prepositional phrase fragments.

In summary, a minimalist corpus-driven approach, beginning with only the existence of word forms, shows that words in English co-occur in highly frequent fixed sequences. These sequences are not complete constituents recognized by traditional theories, but they are readily interpretable in both structural and functional terms.

8.3.2 The interdependence of lexis, grammar, and meaning: Pattern grammar

Many scholars working within a corpus-driven framework have focused on the meaning and use of particular words, arguing that lexis, grammar, and meaning are fundamentally intertwined (e.g., Francis et al. 1996; 1998; Hunston and Francis 1998; 2000; Sinclair 1991a; Stubbs 1993; and Tognini-Bonelli 2001). The best-developed (p. 175) application of corpus-driven research with these goals is the “pattern grammar” reference book series (e.g., Francis et al. 1996; 1998; see also Hunston and Francis 2000).

The pattern grammar studies might actually be considered hybrids, combining corpus-based and corpus-driven methodologies. They are corpus-based in that they assume the existence (and definition) of basic part-of-speech categories and some syntactic constructions, but they 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 … Patterns and lexis are mutually dependent, in that each pattern occurs with a restricted set of lexical items, and each lexical item occurs with a restricted set of patterns. In addition, patterns are closely associated with meaning, firstly because in many cases different senses of words are distinguished by their typical occurrence in different patterns; and secondly because words which share a given pattern tend also to share an aspect of meaning" (Hunston and Francis 2000: 3). Thus, a pattern is a combination of words that "occurs relatively frequently", is "dependent on a particular word choice", and has "a clear meaning associated with it" (Hunston and Francis 2000: 37). 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.

A central claim of this framework is that grammatical patterns have inherent meaning, shared across the set of words that can occur in a pattern. For example, many of the verbs that occur in the grammatical pattern V+ *over* +NP express meanings relating to conflict or disagreement, such as *bicker, disagree, fight, quarrel, quibble*, and *wrangle* (see Hunston and Francis 2000: 43–4); thus it can be argued that the grammatical pattern itself somehow entails this meaning.

The pattern grammar reference books (Francis et al. 1996; 1998) have attempted to provide a comprehensive catalog of the grammatical patterns for verbs, nouns, and adjectives in English. 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. For example, the reference book on grammatical patterns for verbs (Francis et al. 1996) includes over 700 different patterns and catalogs the use of over 4,000 verbs with respect to those patterns. The reference book on grammatical patterns for nouns and adjectives (Francis et al. 1998) is similar in scope, with over 200 patterns used to describe the use of over 8,000 nouns and adjectives.

The pattern grammar reference books do not address some of the stronger theoretical claims that have been associated with the corpus-driven approach. For example, "patterns" are based on analysis of lemmas rather than individual word (p. 176) forms,

and thus the pattern grammar studies provide no support for the general claim that each word form has its own grammar.[1]

The pattern grammar studies also do not support the strong version of the claim that each grammatical pattern has its own meaning. In fact, it is rarely the case that a grammatical frame corresponds to a single meaning domain. However, these studies do provide extensive support for a weaker form of the claim, documenting how the words that occur in a grammatical frame belong to a relatively small set of meaning groups. For example, the adjectives that occur in the grammatical frame **ADJ *in* N** mostly fall into several major meaning groups, such as:

- adjectives that express high interest or participation:
e.g., *absorbed, embroiled, engaged, engrossed, enmeshed, immersed, interested, involved, mixed up, wrapped up*
- adjectives that express a deficit:
e.g., *deficient, lacking, wanting*
- adjectives that express an amount or degree:
e.g., *awash, high, low, poor, rich*
- adjectives that express proficiency or fluency
e.g., fluent, proficient, schooled, skilful, skilled, versed
- adjectives that express that something is covered
e.g., *bathed, clad, clothed, coated, plastered, shrouded, smothered*
(see Francis et al. 1998: 444–51; Hunston and Francis 2000: 75–6).

As noted above, the methodology used for the pattern grammar studies relaxes the strict requirements of corpus-driven methodology. First, predefined grammatical constructs are used in the approach, including basic grammatical classes, phrase types, and even distinctions that require a priori syntactic analysis. In addition, frequency plays only a minor role in the analysis, and some word combinations that occur frequently are not regarded as patterns at all. For example, the nouns followed by complementizer *that* are analyzed as patterns (*e.g., fact, claim, stipulation, expectation, disgust, problem*, etc.), but nouns followed by the relative pronoun *that* do not constitute a pattern, even if the combination is frequent (e.g., *extent, way, thing, questions, evidence, factors + that*). Similarly, prepositions are analyzed for their syntactic function in the sequence noun + preposition, to distinguish between prepositional phrases functioning as adverbials (which do not count as part of any pattern), vs. prepositional phrases that complement the preceding noun (p. 177) (which do constitute a pattern). So, for example, the combinations for the pattern **ADJ *in* N** listed above all include a prepositional phrase that complements the adjective. In contrast, when the prepositional phrase has an adverbial

function, it is analyzed as *not* representing a pattern, even if the combination is frequent. Thus, the following adjectives do not belong to any pattern when they occur in the combination **ADJ *in* N**, even though they occur frequently and represent relatively coherent meaning groups:

> *adamant, firm, resolute, steadfast, unequivocal*
> *loud, vehement, vocal, vociferous*
> (see Hunston and Francis 2000: 76).

Regardless of the specific methodological considerations, the corpus-driven approach as realized in the pattern grammar studies has shown that there are systematic regularities in the associations between grammatical frames, sets of words, and particular meanings, on a much more comprehensive scale than it could have been possible to anticipate before the availability of large corpora and corpus-analysis tools.

8.3.3 The role of frequency in corpus-driven analysis

Surprisingly, one major difference among corpus-driven studies concerns the role of frequency evidence. Nearly every description of the corpus-driven approach includes mention of frequency, as in: (a) the "linguistic categories" are derived "systematically from the recurrent patterns and the frequency distributions that emerge from language in context" (Tognini-Bonelli 2001: 87); (b) in a grammar pattern, "a combination of words occurs relatively frequently" (Hunston and Francis 2000: 37).

In the study of lexical bundles, frequency evidence is primary. This framework can be regarded as the most extreme test of the corpus-driven approach, addressing the question of whether the most commonly occurring sequences of word forms can be interpreted as linguistically significant units. In contrast, frequency is not actually important in pattern grammar studies. On the one hand, frequent word combinations are not included in the pattern analysis if they represent different syntactic constructions, as described in the last section. The combination *satisfaction that* provides another example of this type. When the *that* initiates a complement clause, this combination is one of the realizations of the "happiness" **N *that*** pattern (Francis et al. 1998: 111), as in:

> *One should of course record one's satisfaction that the two leaders got on well together*.

However, it is much more frequent for the combination *satisfaction that* to represent different syntactic constructions, as in:

(p. 178) **(a)** *The satisfaction provided by conformity is in competition with the often more immediate satisfaction that can be provided by crime.*
(b) *He then proved to his own satisfaction that all such endeavours were doomed to failure.*

In (a), the word *that* initiates a relative clause, and in (b), the *that* initiates a verb complement clause controlled by *proved.* Neither of these combinations are analyzed as belonging to a pattern, even though they are more frequent than the combination of *satisfaction* followed by a *that* noun complement clause.
Thus, frequency is not a decisive factor in identifying "patterns", despite the definition that requires that the combination of words in a pattern must occur "relatively frequently". Instead, the criteria that a grammatical pattern must be associated with a particular set of words and have a clear meaning are more decisive (see Hunston and Francis 2000: 67–76).

In fact, some corpus-driven linguists interested in the lexis—grammar interface have overtly argued against the importance of frequency. For example, Sinclair notes that

> some numbers are more important than others. Certainly the distinction between o 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 some 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. Here again, we see the central defining characteristic of corpus-driven research to be the shared goal of identifying new linguistic constructs through inductive analysis of a corpus, regardless of differences in the specific methodological approaches.

(p. 179) 8.3.4 Linguistic “dimensions” of register variation

As discussed in section 8.2 above, corpus research has been used to describe particular linguistic features and their variants, showing how these features vary in their distribution and patterns of use across registers. This relationship can also be approached from the opposite perspective, with a focus on describing the registers rather than describing the use of particular linguistic features.

It turns out, though, that the distribution of individual linguistic features cannot reliably distinguish among registers. There are simply too many different linguistic characteristics to consider, and individual features often have idiosyncratic distributions. Instead, sociolinguistic research has argued that register descriptions must be based on linguistic co-occurrence patterns (see, for example, Ervin-Tripp 1972; Hymes 1974; Brown and Fraser 1979: 38–9; Halliday 1988: 162).

Multi-Dimensional (MD) analysis is a corpus-driven methodological approach that identifies the frequent linguistic co-occurrence patterns in a language, relying on inductive empirical/quantitative analysis (see, for example, Biber 1988; 1995). Frequency plays a central role in the analysis, since each dimension represents a constellation of linguistic features that frequently co-occur in texts. These “dimensions” of variation can be regarded as linguistic constructs not previously recognized by linguistic theory. Thus, although the framework was developed to describe patterns of register variation (rather than the meaning and use of individual words), MD analysis is clearly a corpus-driven methodology in that the linguistic constructs—the “dimensions”—emerge from analysis of linguistic co-occurrence patterns in the corpus.

The set of co-occurring linguistic features that comprise each dimension is identified quantitatively. That is, based on the actual distributions of linguistic features in a large corpus of texts, statistical techniques (specifically factor analysis) are used to identify the sets of linguistic features that frequently co-occur in texts.

The original MD analyses investigated the relations among general spoken and written registers in English, based on analysis of the LOB (Lancaster—Oslo—Bergen) Corpus (15 written registers) and the London—Lund Corpus (six spoken registers). Sixty-seven different linguistic features were analyzed computationally in each text of the corpus. Then, the co-occurrence patterns among those linguistic features were analyzed using

factor analysis, identifying the underlying parameters of variation: the factors or "dimensions". In the 1988 MD analysis, the 67 linguistic features were reduced to seven underlying dimensions. (The technical details of the factor analysis are given in Biber 1988, Chapters 4–5; see also Biber 1995, Chapter 5).

The dimensions are interpreted functionally, based on the assumption that linguistic co-occurrence reflects underlying communicative functions. That is, linguistic features occur together in texts because they serve related communicative functions.

(p. 180) The most important features on Dimensions 1–5 in the 1988 MD analysis are:

Dimension 1: Involved vs. Informational Production

Positive features: mental (private) verbs, *that* complementizer deletion, contractions, present tense verbs, WH-questions, 1st and 2nd person pronouns, pronoun *it*, indefinite pronouns, *do* as pro-verb, demonstrative pronouns, emphatics, hedges, amplifiers, discourse particles, causative subordination, sentence relatives, WH-clauses

Negative features: nouns, long words, prepositions, type/token ratio, attributive adjectives

Dimension 2: Narrative vs. Non-narrative Discourse

Positive features: past tense verbs, 3rd person pronouns, perfect aspect verbs, communication verbs

Negative features: present tense verbs, attributive adjectives

Dimension 3: Situation-dependent vs. Elaborated Reference

Positive features: time adverbials, place adverbials, other adverbs

Negative features: WH-relative clauses (subject gaps, object gaps), phrasal coordination, nominalizations

Dimension 4: Overt Expression of Argumentation

Positive features: prediction modals, necessity modals, possibility modals, suasive verbs, conditional subordination, split auxiliaries

Dimension 5: Abstract/Impersonal Style

Positive features: conjuncts, agentless passives, BY-passives, past participial adverbial clauses, past participial postnominal clauses, other adverbial subordinators

Each dimension can have “positive” and “negative” features. Rather than reflecting importance, positive and negative signs identify two groupings of features that occur in a complementary pattern as part of the same dimension. That is, when the positive features occur together frequently in a text, the negative features are markedly less frequent in that text, and vice versa.

On Dimension 1, the interpretation of the negative features is relatively straightforward. Nouns, word length, prepositional phrases, type/token ratio, and attributive adjectives all reflect an informational focus, a careful integration of information in a text, and precise lexical choice. Text Sample 1 illustrates these co-occurring linguistic characteristics in an academic article:

Text Sample 1. Technical academic prose

Apart from these very general group-related aspects, there are also individual aspects that need to be considered. Empirical data show that similar processes can (p. 181) be guided quite differently by users with different views on the purpose of the communication.

This text sample is typical of written expository prose in its dense integration of information: frequent nouns and long words, with most nouns being modified by attributive adjectives or prepositional phrases (e.g., *general group-related aspects, individual aspects, empirical data, similar processes, users with different views on the purpose of the communication*).

The set of positive features on Dimension 1 is more complex, although all of these features have been associated with interpersonal interaction, a focus on personal stance, and real-time production circumstances. For example, first and second person pronouns, WH-questions, emphatics, amplifiers, and sentence relatives can all be interpreted as reflecting interpersonal interaction and the involved expression of personal stance (feelings and attitudes). Other positive features are associated with the constraints of real time production, resulting in a reduced surface form, a generalized or uncertain presentation of information, and a generally “fragmented” production of text; these include *that*-deletions, contractions, pro-verb DO, the pronominal forms, and final (stranded) prepositions. Text Sample 2 illustrates the use of positive Dimension 1 features in a workplace conversation:

Text Sample 2. Conversation at a reception at work

S□ □ □ □ I□m dying of thirst.

S□ □ □ □ Mm, hmm. Do you need some M & Ms?

S□ □ □ □ Desperately. <laugh> Ooh, thank you. Ooh, you're so generous.

S□ □ □ □ Hey I try.

S□ □ □ □ Let me have my Snapple first. Is that cold-cold ?

S□ □ □ □ I don't know but there should be ice on uh, <unclear>.

S□ □ □ □ I don't want to seem like I don't want to work and I don't want to seem like a stuffed shirt or whatever but I think this is really boring.

S□ □ □ □ I know.

S□ □ □ □ I would like to leave here as early as possible today, go to our rooms, and pick up this thing at eight o'clock in the morning.

S□ □ □ □ Mm, hmm.

Overall, Factor 1 represents a dimension marking interactional, stance-focused, and generalized content (the positive features mentioned earlier) vs. high informational density and precise word choice (the negative features). Two separate communicative parameters seem to be represented here: the primary purpose of the writer/speaker (involved vs. informational), and the production circumstances (those restricted by real-time constraints vs. those enabling careful editing possibilities). Reflecting both of these parameters, the interpretive label “Involved (p. 182) vs. Informational Production” was proposed for the dimension underlying this factor.

The second major step in interpreting a dimension is to consider the similarities and differences among registers with respect to the set of co-occurring linguistic features. To achieve this, *dimension scores* are computed for each text, by summing the individual scores of the features that co-occur on a dimension (see Biber 1988: 93–7). For example, the Dimension 1 score for each text was computed by adding together the frequencies of private verbs, *that*-deletions, contractions, present tense verbs, etc.—the features with positive loadings—and then subtracting the frequencies of nouns, word length, prepositions, etc.—the features with negative loadings.

Once a dimension score is computed for each text, the mean dimension score for each register can be computed. Plots of these mean dimension scores allow linguistic characterization of any given register, comparison of the relations between any two registers, and a fuller functional interpretation of the underlying dimension.

For example, Figure 8.7 plots the mean dimension scores of registers along Dimension 1 from the 1988 MD analysis. The registers with large positive values (such as face-to-face

and telephone conversations), have high frequencies of present tense verbs, private verbs, first and second person pronouns, contractions, etc.—the features with salient positive weights on Dimension 1. At the same time, registers with large positive values have markedly low frequencies of nouns, prepositional phrases, long words, etc.—the features with salient negative weights on Dimension 1. Registers with large negative values (such as academic prose, press reportage and official documents) have the opposite linguistic characteristics: very high frequencies of nouns, prepositional phrases, etc., plus low frequencies of private verbs, contractions, etc.

The relations among registers shown in Figure 8.7 confirm the interpretation of Dimension 1 as distinguishing among texts along a continuum of involved vs. informational production. At the positive extreme, conversations are highly interactive and involved, with the language produced under real-time circumstances. Registers such as public conversations (interviews and panel discussions) are intermediate: they have a relatively informational purpose, but participants interact with one another and are still constrained by real time production. Finally, at the negative extreme, registers such as academic prose are non-interactive but highly informational in purpose, produced under controlled circumstances that permit extensive revision and editing.

Click to view larger

Figure 8.7. **Mean scores of registers along Dimension 1: Involved vs. Informational Production (adapted from Figure 7.1 in** Biber **1988)**

Note: Underlining denotes written registers; capitalization denotes spoken registers; F = 111.9, p <.0001, r^2 = 84.3%.

Figure 8.7 shows that there is a large range of variation among spoken registers with respect to the linguistic features that comprise Dimension 1 (“Involved vs. Informational

Production”). Conversation has extremely large positive Dimension 1 scores; spontaneous speeches and interviews have moderately large positive scores; (p. 183) (p. 184) while prepared speeches and broadcasts have scores around o.o (reflecting a balance of positive and negative linguistic features on this dimension). The written registers similarly show an extensive range of variation along Dimension 1. Expository informational registers, like official documents and academic prose, have very large negative scores; the fiction registers have scores around o.o; while personal letters have a relatively large positive score.

This distribution shows that no single register can be taken as representative of the spoken or written mode. At the extremes, written informational prose is dramatically different from spoken conversation with respect to Dimension 1 scores. But written personal letters are relatively similar to spoken conversation, while spoken prepared speeches share some Dimension 1 characteristics with written fictional registers. Taken together, these Dimension 1 patterns indicate that there is extensive overlap between the spoken and written modes in these linguistic characteristics, while the extremes of each mode (i.e., conversation vs. informational prose) are sharply distinguished from one another.

The overall comparison of speech and writing resulting from the 1988 MD analysis is actually much more complex because six separate dimensions of variation were identified and each of these defines a different set of relations among spoken and written registers. For example, Dimension 2 is interpreted as “Narrative vs. Non-narrative Concerns”. The positive features—past tense verbs, third person pronouns, perfect aspect verbs, communication verbs, and present participial clauses—are associated with past time narration. In contrast, the positive features—present tense verbs and attributive adjectives—have non-narrative communicative functions.

The distribution of registers along Dimension 2, shown in Figure 8.8, further supports its interpretation as Narrative vs. Non-narrative Concerns. All types of fiction have markedly high positive scores, reflecting their emphasis on narrating events. In contrast, registers which are typically more concerned with events currently in progress (e.g., broadcasts) or with building arguments rather than narrating (e.g., academic prose) have negative scores on this dimension. Finally, some registers have scores around 0.0, reflecting a mix of narrative and other features. For example, face-to-face conversation will often switch back and forth between narration of past events and discussion of current interactions.

Click to view larger

Figure 8.8. **Mean scores for registers along Dimension 2: Narrative vs. Non-Narrative Discourse (adapted from Figure 7.2 in** Biber **1988)**

Note: Underlining denotes written registers; capitalization denotes spoken registers; F = 32.3, $p < .0001$, $r^2 = 60.8\%$.

Each of the dimensions in the analysis can be interpreted in a similar way. Overall, the 1988 MD analysis showed that English registers vary along several underlying dimensions associated with different functional considerations, including: interactiveness, involvement and personal stance, production circumstances, informational density, informational elaboration, narrative purposes, situated reference, persuasiveness or argumentation, and impersonal presentation of information. **(p. 185)**

(p. 186) Many studies have applied the 1988 dimensions of variation to study the linguistic characteristics of more specialized registers and discourse domains. For example:However, other MD studies have undertaken new corpus-driven analyses to identify the distinctive sets of co-occurring linguistic features that occur in a particular discourse domain or in a language other than English. The following section surveys some of those studies.

***Present-day registers*:**	***Studies*:**
spoken and written university registers	Biber et al. (2002)
AmE vs. BrE written registers	Biber (1987)
AmE vs. BrE conversational registers	Helt (2001)
biology vs. history student and academic writing	Conrad (1996; 2001)
I-M-R-D sections in medical research articles	Biber and Finegan (1994*b*)
direct mail letters	Connor and Upton (2003)
discourse moves in non-profit grant proposals	Connor and Upton (2004)
oral proficiency interviews	Connor-Linton and Shohamy (2001)
academic lectures	Csomay (2005)
conversation vs. TV dialogue	Quaglio (2009)
female/male conversational style	Rey (2001); Biber and Burges (2000)
author styles	Connor-Linton (2001); Biber and Finegan (1994*a*)

***Historical registers*:**	***Studies*:**
written and speech-based registers; 1650–present	Biber and Finegan (1989; 1997)
medical research articles and scientific research articles; 1650—present	Atkinson (1992; 1996; 1999)
19th-century written registers	Geisler (2002)

8.3.4.1 Comparison of the multi-dimensional patterns across discourse domains and languages

Numerous other studies have undertaken complete MD analyses, using factor analysis to identify the dimensions of variation operating in a particular discourse domain in English, rather than applying the dimensions from the 1988 MD analysis (e.g., Biber 1992; 2001; 2006; 2008; Biber and Jones 2006; Biber et al. 2007; Friginal 2008; 2009; Kanoksilapatham 2007; Crossley and Louwerse 2007; Reppen 2001).

Given that each of these studies is based on a different corpus of texts, representing a different discourse domain, it is reasonable to expect that they would (p. 187) each identify a unique set of dimensions. This expectation is reinforced by the tact that the more recent studies have included additional linguistic features not used in earlier MD studies (e.g., semantic classes of nouns and verbs). However, despite these differences in design and research focus, there are certain striking similarities in the set of dimensions identified by these studies.

Most importantly, in nearly all of these studies, the first dimension identified by the factor analysis is associated with an informational focus vs. a personal focus (personal involvement/stance, interactivity, and/or real-time production features). For example:

Study and dimension	*Corpus*	*Linguistic features defining the dimension*
Biber (2001) Dimension 1	18th-c. written and speech-based registers	prepositions, passives, nouns, long words, past tense verbs vs. 1st and 2nd person pronouns, present tense, possibility and prediction modals, *that*-deletion, mental verbs, emphatics
Biber (2006) Dimension 1	university spoken and written registers	nominalizations, long words, nouns, prepositions, abstract nouns, attributive adjectives, passives, stance noun + *to*-clause, etc. vs. contractions, demonstrative pronouns, *it*, 1st person pronouns, present tense, time advs, *that*-omission, WH-questions, etc.
White (1994) Dimension 1	job interviews	long words, nouns, nominalizations, prepositions, WH-questions, 2nd person pronouns vs. 1st person pronouns, contractions, adverbs, discourse particles, emphatics, etc.
Reppen (2001) Dimension 1	elementary school registers	nouns, long words, nominalizations, passives, attributive adjectives, prepositions vs. initial *and*, time adverbials, 3rd person pronouns
Biber (2008) Dimension 1	conversational text types	long words, nominalizations, prepositions, abstract nouns, relative clauses, attributive adjs. vs. contractions, 1st and 2nd person pronouns, activity verbs

It is perhaps not surprising that Dimension 1 in the original 1988 MD analysis was strongly associated with an informational vs. (inter)personal focus, given that the corpus in that study ranged from spoken conversational texts to written expository texts. For the same reason, it is somewhat predictable that a similar dimension (p. 188) would have emerged from the study of 18th-century written and speech-based registers. It is somewhat more surprising that academic spoken and written registers would be defined by a similar linguistic dimension (and especially surprising that classroom teaching is similar to conversation, and strikingly different from academic writing, in the use of these linguistic features). And it was completely unexpected that a similar oral/literate

dimension—realized by essentially the same set of co-occurring linguistic features—would be fundamentally important in highly restricted discourse domains, including studies of job interviews, elementary school registers, and variations among the different kinds of conversation.

A second parameter found in most MD analyses corresponds to narrative discourse, reflected by the co-occurrence of features like past tense, third person pronouns, perfect aspect, and communication verbs (see, for example, the Biber 2006 study of university registers; Biber 2001 on 18th-century registers; and the Biber 2008 study of conversation text types). In some studies, a similar narrative dimension emerged with additional special characteristics. For example, in Reppen's (2001) study of elementary school registers, “narrative” features like past tense, perfect aspect, and communication verbs co-occurred with once-occurring words and a high type/token ratio; in this corpus, history textbooks rely on a specialized and diverse vocabulary to narrate past events. In the job interview corpus (White 1994), the narrative dimension reflected a fundamental opposition between personal/specific past events and experiences (past tense verbs co-occurring with first person singular pronouns) vs. general practice and expectations (present tense verbs co-occurring with first person plural pronouns). In Biber and Kurjian's (2007) study of web text types, narrative features co-occurred with features of stance and personal involvement on the first dimension, distinguishing personal narrative web pages (e.g., personal blogs) from the various kinds of more informational web pages.

At the same time, most of these studies have identified some dimensions that are unique to the particular discourse domain. For example, the factor analysis in Reppen (1994) identified a dimension of “Other-directed idea justification” in elementary student registers. The features on this dimension include second person pronouns, conditional clauses, and prediction modals; these features commonly co-occur in certain kinds of student writings (e.g., *If you wanted to watch TV a lot you would not get very much done*).

The factor analysis in Biber's (2006) study of university spoken and written registers identified four dimensions. Two of these are similar linguistically and functionally to dimensions found in other MD studies: Dimension 1: “Oral vs. literate discourse”; and Dimension 3: “Narrative orientation”. However, the other two dimensions are specialized to the university discourse domain: Dimension 2 is interpreted as “Procedural vs. content-focused discourse”. The co-occurring “procedural” features include modals, causative verbs, second person pronouns, (p. 189) and verbs of desire + *to*-clause; these features are especially common in classroom management talk, course syllabi, and other institutional writing. The complementary “content-focused” features include rare nouns, rare adjectives, and simple occurrence verbs; these co-occurring features are typical of

textbooks, and especially common in natural science textbooks. Dimension 4, interpreted as “Academic stance”, consists of features like stance adverbials (factual, attitudinal, likelihood) and stance nouns + *that*-clause; classroom teaching and classroom management talk is especially marked on this dimension.

A final example comes from Biber's (2008) MD analysis of conversational text types, which identified a dimension of “stance-focused vs. context-focused discourse”. Stance focused conversational texts were marked by the co-occurrence of *that*-deletions, mental verbs, factual verbs + *that*-clause, likelihood verbs + *that*-clause, likelihood adverbs, etc. In contrast, context-focused texts had high frequencies of nouns and WH-questions, used to inquire about past events or future plans. The text type analysis identified different sets of conversations characterized by one or the other of these two extremes.

In sum, corpus-driven MD studies of English registers have uncovered both surprising similarities and notable differences in the underlying dimensions of variation. Two parameters seem to be fundamentally important, regardless of the discourse domain: a dimension associated with informational focus vs. (inter)personal focus, and a dimension associated with narrative discourse. At the same time, these MD studies have uncovered dimensions particular to the communicative functions and priorities of each different domain of use.

These same general patterns have emerged from MD studies of languages other than English, including Nukulaelae Tuvaluan (Besnier 1988); Korean (Kim and Biber 1994); Somali (Biber and Hared 1992; 1994); Taiwanese (Jang 1998); Spanish (Biber et al. 2006; Biber and Tracy-Ventura 2007; Parodi 2007); Czech (Kodytek 2008), and Dagbani (Purvis 2008). Taken together, these studies provide the first comprehensive investigations of register variation in non-western languages.

Biber (1995) synthesizes several of these studies to investigate the extent to which the underlying dimensions of variation and the relations among registers are configured in similar ways across languages. These languages show striking similarities in their basic patterns of register variation, as reflected by:

- the co-occurring linguistic features that define the dimensions of variation in each language;
- the functional considerations represented by those dimensions;
- the linguistic/functional relations among analogous registers.

For example, similar to the full MD analyses of English, these MD studies have all identified dimensions associated with informational vs. (inter)personal purposes, and with narrative discourse.

(p. 190) At the same time, each of these MD analyses have identified dimensions that are unique to a language, reflecting the particular communicative priorities of that language and culture. For example, the MD analysis of Somali identified a dimension interpreted as "Distanced, directive interaction", represented by optative clauses, first and second person pronouns, directional preverbal particles, and other case particles. Only one register is especially marked for the frequent use of these co-occurring features in Somali: personal letters. This dimension reflects the particular communicative priorities of personal letters in Somali, which are typically interactive as well as explicitly directive.

The cross-linguistic comparisons further show that languages as diverse as English and Somali have undergone similar patterns of historical evolution following the introduction of written registers. For example, specialist written registers in both languages have evolved over time to styles with an increasingly dense use of noun phrase modification. Historical shifts in the use of dependent clauses is also surprising: in both languages, certain types of clausal embedding—especially complement clauses—turn out to be associated with spoken registers rather than written registers.

These synchronic and diachronic similarities raise the possibility of universale of register variation. Synchronically, such universals reflect the operation of underlying form/function associations tied to basic aspects of human communication; and diachronically, such universals relate to the historical development of written registers in response to the pressures of modernization and language adaptation.

8.4 Conclusion

The present chapter has illustrated how corpus analysis contributes to the description of language use, in many cases allowing us to think about language patterns in fundamentally new ways. Corpus-based analyses are the most traditional, employing the grammatical categories recognized by other linguistic theories but investigating their patterns of variation and use empirically. Such analyses have shown repeatedly that our intuitions about the patterns of use are often inaccurate, although the patterns themselves are highly systematic and explainable in functional terms.

Corpus-driven approaches are even more innovative, using corpus analysis to uncover linguistic constructs that are not recognized by traditional linguistic theories. Here again, corpus analyses have uncovered strong, systematic patterns of use, (p. 191) but even in this case the underlying constructs had not been anticipated by earlier theoretical frameworks.

In sum, corpus investigations show that our intuitions as linguists are not adequate for the task of identifying and characterizing linguistic phenomena relating to language use. Rather, corpus analysis has shown that language use is patterned much more extensively, and in much more complex ways, than previously anticipated. (p. 192)

Notes:

(1) Other studies that advocate this position have been based on a few selected case studies (e.g., Sinclair 1991*b* on *eye* vs. *eyes;* Tognini-Bonelli and Elena 2001: 92–8 on *facing vs. faced*, and *saper* vs. *sapere* in Italian). These case studies clearly show that word forms belonging to the same lemma do sometimes have their own distinct grammar and meaning. However, no empirical study to date has investigated the extent to which this situation holds across the full set of word forms and lemmas in a language. (In contrast, the pattern grammar reference books seem to implicitly suggest that most inflected word forms that belong to a single lemma "pattern" in similar ways.)

Douglas Biber

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). His publications include books on register variation and corpus linguistics published by Cambridge University Press (1988, 1995, 1998, to appear), the co-authored *Longman Grammar of Spoken and Written English* (1999), and more recent studies of language use in university settings and discourse structure investigated from a corpus perspective (both published by Benjamins: 2006 and 2007).