Corpus linguistics
Investigating language structure and use

DOUGLAS BIBER
Northern Arizona University

SUSAN CONRAD
Iowa State University

RANDI REPPEN
Northern Arizona University

CAMBRIDGE
UNIVERSITY PRESS

Contents

1 Introduction
Goals and methods of the corpus-based approach

1.1 Studying language: structure and use

Studies of language can be divided into two main areas: studies of structure and studies of use. Traditionally, linguistic analyses have emphasized structure – identifying the structural units and classes of a language (e.g., morphemes, words, phrases, grammatical classes) and describing how smaller units can be combined to form larger grammatical units (e.g., how words can be combined to form phrases, phrases can be combined to form clauses, etc.).

A different perspective – which is the focus of this book – is to emphasize language use. From this perspective, we can investigate how speakers and writers exploit the resources of their language. Rather than looking at what is theoretically possible in a language, we study the actual language used in naturally occurring texts.

Many studies of language use focus on a particular linguistic structure, investigating the ways in which seemingly similar structures occur in different contexts and serve different functions. For example, in English *that*-verb-complement clauses and *to*-verb-complement clauses are similar in their structural characteristics and can be similar in meaning, as in sentences such as:

(1) I hope that I can go.
(2) I hope to go.

In addition, *that*-clauses can occur with the *that* omitted:

(3) I hope I can go.

A structural analysis would describe the grammatical similarities and differences among these three sentences. All three options are equally grammatical ways to complete the meaning of the verb. However, an analysis of language use goes beyond traditional grammatical description to ask why the language should have multiple structures that are so similar in their meaning and grammatical function.

Answers to this question should consider a range of factors. For example, do spoken varieties versus written varieties have different preferences for one of the forms over others? Are the forms usually used with different verbs? Are the forms used preferentially for different specialized meanings? These are some of the kinds of questions that can be addressed in studies of use. In fact, in the illustrative analyses presented in Chapters 3 and 4, you will see that there are strong patterns in the preferred use of these different structures.

In addition to analyzing the language use patterns for a linguistic structure, studies of use can focus on the language of a text or a group of speakers/writers. For example, interest in an individual author's style or in the language used by different social groups has been a common motivation for studies of use, considering questions such as: How does the language used by a particular author compare to the language used by his contemporaries? How does the language used by women differ from the language used by men?

Equally important are investigations comparing the language of different texts or groups of texts. Many times each day we use different varieties of language as we participate in different situations – from talking to a family member, to reading a newspaper, to writing a letter to a friend, to reading an academic article. The varieties of language that we use in different situations are referred to as *registers*, and describing the characteristics of these registers is an important area of study. However, it is also a complex one because many different grammatical and lexical choices come into play. How can we find the patterns in the language used in conversation, newspapers, academic prose, personal letters, etc.? How can we characterize the language used in these different varieties? Questions such as these are also an important aspect of studies of use, and are addressed in Part II of this book.

For all these studies of use, analysts attempt to uncover typical patterns rather than making judgements of grammaticality. There are two central research goals in such analyses of use: (1) assessing the extent to which a pattern is found, and (2) analyzing the contextual factors that influence variability. For example, in an analysis of *that*-complement clauses versus *to*-clauses, we would want to consider whether speakers have a preference for one kind of clause and writers have a preference for the other kind of clause. Further, we would want to consider a range of contextual factors, such as the typical verbs that each clause type is used with.

Finding patterns of use and analyzing contextual factors can present difficult methodological challenges. Because we are looking for typical patterns, analyses cannot rely on intuitions or anecdotal evidence. In many cases, humans tend to notice unusual occurrences more than typical occurrences, and therefore conclusions based on intuition can be unreliable. Furthermore, we need to analyze a large amount of language collected from many speakers, to make sure that we are not basing conclusions on a few speakers' idiosyncrasies. However, with a large amount of language, it is time-consuming to carry out the analyses and difficult to keep track of multiple contextual factors. If you wanted to compare the language used in conversation and academic articles, for example, imagine how difficult it would be to keep track of even twenty different linguistic structures in ten texts from each register – let alone figure out the ways that these structures are interrelated with a range of contextual factors.

Because of these difficulties, until recently many investigations of language use were either unfeasible or simply impossible. The corpus-based approach, however, provides a means of handling large amounts of language and keeping track of many contextual factors at the same time. It therefore has opened the way to a multitude of new investigations of language use.

1.2 What is the corpus-based approach?

At this point, you might be wondering: what actually is the corpus-based approach, and what makes it different from other analytical approaches in linguistics? The following sections address

these questions. We begin by identifying the essential characteristics of corpus-based analyses in Section 1.2.1. Underlying these characteristics is a new perspective on language use: studying the use of language characteristics by considering the relevant "association patterns." This notion, which forms the basis for all subsequent analyses in the book, is introduced in Section 1.2.2. Association patterns represent quantitative relations, measuring the extent to which features and variants are associated with contextual factors. However, functional (qualitative) interpretation is also an essential step in any corpus-based analysis, and so we discuss the relationship between quantitative and qualitative techniques in Section 1.2.3. Finally, in Sections 1.2.4 and 1.2.5, we turn to a comparison of the corpus-based approach with other analytical approaches in linguistics, and summarize the research areas that can be studied using this approach.

1.2.1 The characteristics of corpus-based analyses

The essential characteristics of corpus-based analysis are:

- 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;
- it makes extensive use of computers for analysis, using both automatic and interactive techniques;
- it depends on both quantitative and qualitative analytical techniques.

Taken together, these characteristics result in a scope and reliability of analysis not otherwise possible. Several of the advantages of the corpus-based approach come from the use of computers. Computers make it possible to identify and analyze complex patterns of language use, allowing the storage and analysis of a larger database of natural language than could be dealt with by hand. Furthermore, computers provide consistent, reliable analyses – they don't change their mind or become tired during an analysis. Computers can also be used interactively, allowing the human analyst to make difficult linguistic judgements while the computer takes care of record-keeping.

Finally, it is important to note that corpus-based analyses must go beyond simple counts of linguistic features. That is, it is essential to include qualitative, functional interpretations of quantitative patterns. In each chapter of this book, you will find that a great deal of space is devoted to explanation, exemplification, and interpretation of the patterns found in quantitative analyses. The goal of corpus-based investigations is not simply to report quantitative findings, but to explore the importance of these findings for learning about the patterns of language use.

1.2.2 Association patterns in language use

Many early studies in corpus linguistics simply counted the occurrence of linguistic items. For instance, some lexical studies compared the frequency of particular words, or of two-letter, three-letter, and four-letter words. Some grammatical studies counted the frequency of nouns, verbs, and adjectives. Studies of this type can be useful in providing reference materials (such as identifying the fifty most common words) or for providing simple stylistic indicators (such as the relative frequencies of nouns and verbs in a text).

However, a representative corpus, if properly exploited, can provide many additional kinds of information about language use. In particular, a corpus-based approach allows researchers to identify and analyze complex "association patterns": the systematic ways in which linguistic features are used in association with other linguistic and non-linguistic features.

There are two main kinds of research question that can be investigated in terms of association patterns, as shown in Table 1.1. The first is to focus on the use of a linguistic feature – either a lexical item or a grammatical construction (Part A of Table 1.1); the second is to focus on the characteristics of texts or varieties (Part B of Table 1.1).

Linguistic analyses have traditionally focused on a particular linguistic feature, either a word or grammatical construction. However, the use of such features can be further investigated by considering their systematic associations with other features. Two main kinds of associations are important here: linguistic associations and non-linguistic associations.

Table 1.1 *Association patterns in language use*

A.	Investigating the use of a linguistic feature (lexical or grammatical)
	(i) Linguistic associations of the feature
	– lexical associations (associations with particular words)
	– grammatical associations (associations with particular grammatical constructions)
	(ii) Non-linguistic associations of the feature
	– distribution across registers
	– distribution across dialects
	– distribution across time periods
B.	Investigating varieties or texts (e.g., registers, dialects, historical periods)
	(i) Linguistic association patterns
	– individual linguistic features or classes of features
	– co-occurrence patterns of linguistic features

Linguistic associations fall into two major categories:

1. lexical associations – investigating how the linguistic feature is systematically associated with particular words;
2. grammatical associations – investigating how the linguistic feature is systematically associated with grammatical features in the immediate context.

Lexical associations are illustrated in Chapter 2 through an analysis of the words *big*, *large*, and *great*. Specifically, that analysis considers the collocates of these three words – that is, the words that tend to co-occur with each target word. For example, *big* commonly co-occurs with *toe*, while *large* commonly co-occurs with *number*. Although these three words are nearly synonymous in isolation, the analysis shows that they tend to be used with very different kinds of words. Thus, this analysis looks at "lexical–lexical" association patterns and finds them to be quite different for each of these three words.

In Chapter 4, on the other hand, we investigate "lexical–grammatical" associations. For example, we compare the nearly synonymous adjectives *small* and *little*, showing how they have very different grammatical associations with attributive versus

predicative positions (e.g., the *small* boy versus the boy is *small*). The opposite type of research question – focusing on a grammatical feature and considering its lexical associations – is also exemplified in the book. For instance, in Chapter 4 we compare the verbs that are most commonly used with *that*-clauses versus *to*-clauses (such as *think* co-occurring with *that*-clauses versus *want* commonly co-occurring with *to*-clauses).

In addition to its linguistic associations, the use of a linguistic feature can be studied in terms of its non-linguistic associations. Three major factors are relevant here: how a lexical item or grammatical construction is distributed differentially across 1. varieties defined by situation (registers), 2. varieties defined by social group (dialects), or 3. periods of time. For example, Chapter 3 includes an investigation of how nominalizations are distributed differently across academic prose and conversation – an example of the association between a grammatical feature (nominalizations) and non-linguistic feature (register).

It is important to realize that linguistic and non-linguistic association patterns are not independent. Rather, they interact. Thus, most sample analyses in this book include both kinds of association patterns. For example, when we consider lexical–lexical associations for *big*, *large*, and *great*, we also consider their distributions across different registers.

Instead of focusing on particular linguistic features, it is also possible to describe the characteristics of texts or varieties in terms of association patterns (Part B of Table 1.1). In this case, corpus-based studies attempt to characterize registers, dialects, styles, or individual literary works in terms of their linguistic association patterns. These linguistic associations can be either individual features or classes of features. In Chapter 6, for example, we characterize different spoken and written registers with respect to their use of dependent clauses.

However, to characterize varieties more thoroughly, another kind of linguistic association pattern is important: the ways in which groups of linguistic features commonly co-occur in texts. For example, nouns, prepositions, long words, and attributive adjectives tend to co-occur in certain registers. Why is this so? What function do these features share? What other features tend

to occur in texts when these features are rare? These and related questions are addressed in the second half of this book.

Though many different kinds of association patterns can be investigated with corpus-based studies, all of these patterns share an important characteristic: they represent continuous relationships. That is, the patterns are not absolute statements about what always happens or never happens in language use; rather, these patterns occur to differing extents. We might think of certain patterns as very common or very rare – but what does "common" or "rare" signify? Making comparisons between association patterns requires a more precise characterization of the extent to which different patterns exist – that is, quantitative measures. The next section discusses the use of quantitative analyses in corpus-based research, as well as the complementary role of qualitative, functional interpretation.

1.2.3 The role of quantitative analyses and functional interpretations

In the last section we reviewed different kinds of association patterns that can be investigated in corpus-based studies and noted that these relationships are continuous constructs. Therefore, quantitative techniques are essential for corpus-based studies. For example, if you wanted to compare the language use patterns for the words *big* and *large*, you would need to know how many times each word occurs in the corpus, how many different words commonly co-occur with each of these adjectives (the collocations), and how common each of those collocations is. These are all quantitative measurements.

In all the sample analyses in this book, you will find quantitative analyses. For many of the examples, particularly in the early chapters of the book, we present only frequency data – how often a certain pattern occurs relative to other patterns. In many cases strong patterns can be observed directly from these frequencies, and in order to keep examples straightforward and accessible to a wide audience, we do not use statistical procedures in these cases.

In some of the examples in later chapters of the book, however, statistical procedures are important for investigating complex

association patterns. For some analyses, tests of statistical "significance" are also included. Significance tests show how likely it is that quantitative results could have occurred by chance, and thus they should always be reported in research articles describing a corpus-based study. However, it is not our purpose here to teach you how to carry out statistical tests. We do provide conceptual introductions to the statistical procedures used in our example analyses, and we provide some methodological details in the methodology boxes included in Part IV. The discussion in this book will allow you to understand the purpose and importance of the statistical procedures that we have used, but we have not set out to comprehensively discuss statistical techniques. You should consult a statistics textbook for more complete coverage of that kind.

In addition, as you read the sample analyses in this book you will find much more than quantitative and statistical findings. As noted above, a crucial part of the corpus-based approach is going beyond the quantitative patterns to propose functional interpretations explaining why the patterns exist. As a result, a large amount of effort in corpus-based studies is devoted to explaining and exemplifying quantitative patterns. In a textbook this size it is not possible to provide a full functional interpretation of every analysis. However, we do consistently outline the major aspects of such interpretations, emphasizing the importance of this step in all corpus-based analyses.

1.2.4 The corpus-based approach compared to other approaches in linguistics

So far in this introduction, we have been emphasizing the distinctive features of the corpus-based approach. In particular, we have emphasized its strengths in investigating language use, as opposed to traditional studies of language structure. We have noted that comprehensive studies of use cannot rely on intuition, anecdotal evidence, or small samples; they rather require empirical analysis of large databases of authentic texts, as in the corpus-based approach.

However, corpus-based analysis should be seen as a complementary approach to more traditional approaches, rather than as