Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers
Gena R. Bennett
http://www.press.umich.edu/titleDetailDesc.do?id=371534
Michigan ELT, 2010

PART 1

AN INTRODUCTION TO CORPUS LINGUISTICS

The principles of corpus linguistics have been around for almost a century. Lexicographers, or dictionary makers, have been collecting examples of language in use to help accurately define words since at least the late 19th century. Before computers, these examples of language were essentially collected on small slips of paper and organized in pigeon holes. The advent of computers led to the creation of what we consider to be modern-day corpora. The first computer-based corpus, the Brown corpus, was created in 1961 and comprised about 1 million words. Today, generalized corpora are hundreds of millions of words in size, and corpus linguistics is making outstanding contributions to the fields of second language research and teaching.

WHAT IS CORPUS LINGUISTICS?

So what exactly is corpus linguistics? Corpus linguistics approaches the study of language in use through corpora (singular: *corpus*). A corpus is a large, principled collection of naturally occurring examples of language stored electronically. In short, corpus linguistics serves to answer two fundamental research questions:

1. What particular patterns are associated with lexical or grammatical features?
2. How do these patterns differ within varieties and registers?

Many notable scholars, have, of course, contributed to the development of modern-day corpus linguistics: Leech, Biber, Johansson, Francis, Hunston, Conrad, and McCarthy, to name just a few. These scholars have made substantial contributions to corpus linguistics, both past and present. Many corpus linguists, however, consider John Sinclair to be one of, if not the most, influential scholar of modern-day corpus linguistics. Sinclair detected that a word in and of itself does not carry meaning, but that meaning is often made through several words in a sequence (Sinclair, 1991). This is the idea that forms the backbone of corpus linguistics.

WHAT CORPUS LINGUISTICS IS NOT

It's important to not only understand what corpus linguistics is, but also what corpus linguistics is <u>not</u>. Corpus linguistics is not

- able to provide negative evidence
- able to explain why
- able to provide all possible language at one time.

Corpus linguistics is not able to provide negative evidence. This means a corpus can't tell us what's possible or correct or not possible or incorrect in language; it can only tell us what is or is not present in the corpus. Many instructors mistakenly believe that if a corpus does not present all manners to express a certain idea, then the corpus is altogether faulty. Instead, instructors should believe that if a corpus does not present a particular manner to express a certain idea, then perhaps that manner is not very common in the register represented by the corpus.

Corpus linguistics is not able to explain why something is the way it is, only tell us what is. To find out why, we, as users of language, use our intuition.

Corpus linguistics is not able to provide all possible language at one time. By definition, a corpus should be principled: "a large, *principled* collection of naturally occurring texts. . .," meaning that the language that goes into a corpus isn't random, but planned. However, no matter how planned, principled, or large a corpus is, it cannot be a representative of all language. In other words, even in a corpus that contains one billon words, such as the Cambridge International Corpus (CIC), all instances of use of a language may not be present.

Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers
Gena R. Bennett
http://www.press.umich.edu/titleDetailDesc.do?id=371534
Michigan ELT, 2010

Chapter 1

Principles of Corpus Linguistics

QUESTIONS WE CAN ANSWER WITH CORPORA

Broadly, corpus linguistics looks to see what patterns are associated with lexical and grammatical features. Searching corpora provides answers to questions like these:

- What are the most frequent words and phrases in English?
- What are the differences between spoken and written English?
- What tenses do people use most frequently?
- What prepositions follow particular verbs?
- How do people use words like *can, may*, and *might*?
- Which words are used in more formal situations and which are used in more informal ones?
- How often do people use idiomatic expressions?
- How many words must a learner know to participate in everyday conversation?
- How many different words do native speakers generally use in conversation? (McCarthy, 2004, pp. 1–2)

For the most part, these questions don't look particularly revolutionary. We already know the answers to a lot of them. We teach the ideas contained within many of these questions every day. We can open up almost any grammar, vocabulary, conversation, or writing textbook and find the answers. Even better, we can apply our expert-user intuition to find the answers. We're intimately connected to the language; after all, we speak it every day, right? An exercise may help here. For example, O'Keeffe, McCarthy, and Carter (2007, p. 32) studied a frequency list from a 10 million–word corpus and discovered that the 2,000 most frequent words in the corpus accounted for 80 percent of all the

A **frequency list** displays the words occurring in a corpus along with the number of times each word appears.

Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers
Gena R. Bennett
http://www.press.umich.edu/titleDetailDesc.do?id=371534
Michigan ELT, 2010

words present. A mere 2 percent of the words were used repeatedly to account for 8 million words.

For example, degree adverbs demonstrate the extent of a particular feature, such as *thoroughly* in the sentence, *Her chocolate cake is thoroughly delicious*. Keep this in mind, and think for a moment about these questions.

- What are some common adverbs of degree? Think of at least four.
- Give examples of ways you would use these adverbs.
- Which adverbs do you think are used more often in speaking?
- Which adverbs do you think are used more often in writing?
- Which adverbs do you think are used more often overall?

You may have thought of these, among others:

- ***very***—*My sister is very intelligent.*
- ***really***—*Listening to an in-class lecture can be really difficult.*
- ***exactly***—*Sue always knows exactly what I'm thinking.*
- ***quite***—*Frederick appeared quite surprised by the low mark on his project.*
- ***completely***—*The surprise birthday party was completely unexpected.*
- ***too***—*Working full time and going to school full time is too demanding for my schedule.*

From this list of adverbs, we might think that *really* is used more in speaking and *quite* is used more in writing. Perhaps *very* is used most frequently overall.

The exercise used multiple adverbs of degree: where they're used, the frequency of use, and some examples of use. This information seems like sufficient material for a lesson, and most teachers would feel comfortable presenting this information in class.

Corpora can give us information like frequency, register, and how language is used, ideas identified in the adverbs of degree exercise.

Table 1.1 shows the frequency results per million (rounded to the nearest one) from the Corpus of Contemporary American English (COCA). (See Appendix 1 for

Because corpora don't contain the same number of words, we can't use a simple frequency count to see in which corpus a word is more common. For example, *very* occurs in the spoken portion of the Corpus of Contemporary American English (COCA) 195,000 times and in the written portion of the COCA 198,000 times; from looking only at the simple frequency count, we might conclude that *very* is used only slightly more in written language. But, because the written portion of the COCA is much larger than the spoken portion, we can only get an accurate comparison by calculating how many times *very* occurs per million words. This is the **normed count**. The normed counts in Table 1.1 show that for every million words in the spoken portion of the COCA, *very* appears 2,543 times; for every million words in the written portion, *very* only appears 673 times. This allows us to see that, in fact, *very* is used significantly more frequently in the spoken portion of the corpus than in the written portion of the corpus.

more information on this and other corpora. COCA will also be discussed in Chapter 4.) The numbers in the Speak column indicate how many times the adverbs *very, really, exactly, quite, completely, too,* and *thoroughly* are used in the spoken portion of the COCA. The numbers in the Write column indicate how many times the adverbs are used in the written portion of the COCA, and the numbers in the TOTAL columns indicate how many times the adverbs are used overall.

Very is the most frequently used adverb overall in the COCA, and is especially frequent in spoken language. *Really* is the second most frequent adverb in speaking and overall, while *too* is the most frequent adverb used in writing. Note that *too* and *completely* are used almost the same in speaking and as in writing. With the exception of *thoroughly*, and to a significantly lesser degree *too*, these adverbs of degree are used more frequently in spoken language than written language.

Visit www.americancorpus.org/ to complete your own search on **adverbs of degree.**

So what does all this mean? These data present us with opportunities to show students more accurately how to use language. When teaching adverbs of degree, based on the information from this corpus, it would be prudent to emphasize the following:

- Focus attention on *really* and *very* because they are the most commonly used adverbs of degree, and students will likely encounter them often.
- Point out that *too* and *completely* are different from *really* and *very* because they are used almost equally in speaking and writing.
- Show how *thoroughly* is used differently because it appears more frequently in writing than in speaking (the only adverb here to significantly do so).

Table 1.1
Frequency Results Per Million of Adverbs of Degree in COCA

Word	Speak	Write	Total
very	2,543	673	3,216
really	1,637	392	2,029
exactly	271	93	364
quite	267	150	417
completely	87	78	165
too	656	699	1,355
thoroughly	7	18	25
Total	5,468	2,103	7571

Source: Corpus of Contemporary American English

- Use more listening and speaking activities to teach adverbs of degree because these particular adverbs are used more than two times more in speaking than writing.

In a nutshell, corpus linguistics allows us to see how language is used today and how that language is used in different contexts, enabling us to teach language more effectively.

THE CORPUS APPROACH

Is corpus linguistics a methodology? Is it theory? Most corpus linguists are not willing to answer that question in such terms, but when analyzing language using corpora, there is a "method" to employ.

The Corpus Approach (Biber, Conrad, & Reppen, 1998, p. 4) is comprised of four major characteristics:

1. It is empirical, analyzing the actual patterns of language use in natural texts.
2. It utilizes a large and principled collection of natural texts as the basis for analysis.
3. It makes extensive use of computers for analysis.
4. It depends on both quantitative and qualitative analytical techniques.

1. The Corpus Approach is empirical, analyzing the actual patterns of language use in natural texts.

The key to this characteristic of the Corpus Approach is authentic language. The idea that corpora are principled has been mentioned but not what language a corpus is comprised of. Corpora are composed from textbooks, fiction, nonfiction, magazines, academic papers, world literature, newspapers, telephone conversations at home or work, cell phone conversations, business meetings, class lectures, radio broadcasts, and TV shows, among other communication acts. In short, any real-life situation in which any linguistic communication takes place can form a corpus.

2. The Corpus Approach utilizes a large and principled collection of naturally occurring texts as the basis for analysis.

This characteristic of the Corpus Approach refers to the corpus itself. You may work with a written corpus, a spoken corpus, an academic spoken corpus, etc.

3. The Corpus Approach makes extensive use of computers for analysis.

Not only do computers hold corpora, they help analyze the language in a corpus. A corpus is accessed and analyzed by a concordancing program. In short, you can't effectively utilize corpora, or employ the corpus approach, without a computer.

4. The Corpus Approach depends on both quantitative and qualitative analytical techniques.

This characteristic of the corpus approach highlights the importance of our intuition as expert users of a language. We take the quantitative results generated from the corpus and then analyze them qualitatively to find significance. Table 1.1 shows quantitative results. Qualitatively analyzing the results would involve examining the adverbs of degree in use to understand situations the adverbs are used in. This is how we answer the question *Why*?

TARGET FEATURES

Although intuition may not always be reliable for drawing conclusions about language in general, it does often answer the question *Why*? Intuition is often useful for helping us form queries for a corpus. Many of the questions that corpora answers fall into certain areas of language teaching, such as phraseology, lexicogrammar, registers, English for Specific Purposes (ESP), nuances of language, and appropriate syllabus design.

Phraseology

Phraseology is the study of phrases. Phraseology is a central element of corpus linguistics: Sinclair (1991) determined that the meaning of a word is found through several words in a sequence, through phrases. Phraseology includes the study of collocations, lexical bundles, and language occurring in preferred sequences.

Collocation

The most prominent way of studying phrases is through collocation. Collocation is the statistical tendency of words to co-occur. This means that when one word is used, there is a high statistical probability that a certain word or words will occur alongside of it. For example, look at the noun form of the word *deal*. The words *big, good,* and *great* are collocations of *deal* as a noun, meaning that when we use *deal* as a noun, we often refer to a *big deal*, a *good deal*, and/or a *great deal*. From studying collocations, we know that there is a tendency for each collocate of a word to be asso-

ciated with a single sense of that word. We can see this looking at the phrases using *deal—big deal, good deal, great deal*: a *big deal* is usually an event or situation that has significant meaning; a *good deal* generally refers to a bargain; a *great deal* often refers to a quantity. Studying collocations provides a deeper understanding of the meaning and use of a word, such as *deal*, than simply studying a word alone.

Collocations can also help us better understand particular words used in a certain phrase. Kennedy (1991) studied *between* and *through*, something many language textbooks have difficulty distinguishing the use of. By studying the collocations of the two words, Kennedy found that *between* is usually used after nouns like *differences, distinction, agreement*, and *meeting*, whereas *through* is more frequently found after verbs such as *go, pass, run,* and *fall* (1991, p. 107). (Chapter 3 of this book presents an activity dealing with collocation.)

Lexical Bundles

Phraseology also looks at variation in somewhat fixed phrases, which are often referred to as lexical bundles. Biber, Johansson, Leech, Conrad, and Finegan (1999, p. 990) define a lexical bundle as a recurring sequence of three or more words. In conversation, "Do you want me to" and "I don't know what" are among the most common lexical bundles (Biber et al., 1999, p. 994). It is important to understand that lexical bundles are different from idioms. Idioms have a meaning not derivable from their parts, unlike lexical bundles, which do. Also, lexical bundles are not complete phrases. Most important, lexical bundles are statistically defined (identified by their overwhelming co-occurrence), and idioms are not.

One type of lexical bundle is a frame. A frame has set words around a variable word or words. One example of the use of frames is the expression of future time. In the Corpus of Contemporary American English, multiple words are used to express future time using the frame *is…to*: *is going to, is likely to, is expected to, is supposed to, is about to, is due to*. *Is* and *to* are the set words of the frame that surround the variables like *going likely, expected,* or *about*.

Preferred Sequences

Phraseology also includes the study of preferred sequences of words. Look at the adjectives *interested* and *interesting*. Hunston (2002, pp. 9–11) explains that learners often confuse these two words, and explanations of their different meanings do not usually help students use the words correctly; looking at the phrases *someone is interested in something*, *an interesting thing*, *what is interesting*, and *it is interesting to see*, can give students the ability to use the individual words correctly by providing an established pattern of use for each word.

Only through corpus study can we find the details of phraseology—collocations, lexical bundles, and language occurring in preferred sequences.

Lexicogrammar

Another area of language teaching that corpus linguistics addresses is lexicogrammar. Lexicogrammar is Sinclair's (1991) idea that there is no difference between lexis and grammar, or that lexis and grammar are so closely intertwined that they cannot be productively studied separately. Certain lexical items fall in certain patterns and certain patterns contain certain lexical items.

To see the **most common patterns** for verbs, nouns, or adjectives, visit http://candle.cs.nthu.edu.tw/collocation/webform2.aspx?funcID=9.

An example of the idea of lexicogrammar includes certain words (lexicon) associated with certain verbs tenses (grammar): *know, matter*, and *suppose* occur more than 80 percent of the time in the present tense while *smile, reply,* and *pause* occur more than 80 percent in the past tense (Biber et al., 1999, p. 459). We can also find that some verbs are used most frequently in certain clauses: *know* and *think* are associated with *that*-complement clauses (Biber et al., 1999, p. 661), while the verbs *like, want,* and *seem* are associated with *to*-complement clauses (Biber et al., 1999, pp. 699).

Hunston and Francis (2000, p. 1–2) offer us a more detailed example.

> ❶ Philosophy is different from many other disciplines ❷ in that learning about it is as much a matter of developing skills (in reasoning and argument) as it is a matter of learning a body of information. ❸ In this sense there are no definitive 'answers' to many philosophical problems: ❹ becoming a philosopher is a matter of becoming able to reason coherently and relevantly about philosophical issues. ❺ Consequently, valuable contact time with lecturers is best spent actually 'doing philosophy,' ❻ and that means actively thinking and talking about it.

The word *matter* appears three times in the paragraph. What is happening in each usage?

a matter of developing skills

a matter of learning a body of information

a matter of becoming able to reason coherently and relevantly

a matter of + *-ing*.

In this sense or use of *matter*, it's much more productive to teach the pattern *a matter of* + *-ing* rather than to focus on the single lexical word *matter*.

Register

The third area of language teaching that corpus linguistics addresses is register. Register is defined as situation of use.[1] We use different language with different audiences—our parents, colleagues, or children—at different times and for different reasons. Register can be broadly defined—spoken versus written—or more narrowly defined—conversation versus news or even separate parts of a research paper.

Corpus linguistics addresses language teaching through the study of registers by illustrating the various phraseology and lexicogrammar used from register to register. For example, 90 percent of lexical bundles in conversation are declarative or interrogative clauses (Biber et al., 1999, p. 999); pronouns are used slightly more in conversation than nouns, but nouns are used significantly more than pronouns in fiction, news, and academic writing (Biber et al., 1999, p. 235); past tense is used more in writing and present tense more in conversation (Biber et al., 1999, p. 456). In different registers, corpora show us differences of use in language such as word frequency, word meaning and use, and grammatical frequency.

ESP (English for Specific Purposes)

ESP is probably one of the most obvious and pointed applications of corpus linguistics. The areas of register, lexicogrammar, and phraseology can all be applied to specific purposes.

Visit the official **AWL** website: www.victoria.ac.nz/lals/staff/averil-coxhead/awl/.

The Academic Word List (AWL) (Coxhead, 2000) is a well-known example of using corpus linguistics to address ESP, in this case, academic purposes. By investigating a corpus comprised of academic language, Coxhead was able to pinpoint the most frequent vocabulary words used in academic texts; she then made the list available for instructors to help students focus their vocabulary study. A project is also underway to further that study by investigating the phraseology and lexicogrammatical patterns of the top words on the AWL (Byrd, 2007).

Like the corpus created using academic texts to compile the AWL, corpora can be created and investigated for a myriad of purposes. Right now corpora exist for nurses and health care professionals, air traffic controllers, and switchboard operators, just to name a few. Iowa State University has used a corpus of research articles for each major of its graduate students in order to help them write research articles in their designated field (Cortes, 2007).

The **BLC (Business Letters Corpus)** is an ESP corpus. You can search the BLC to explore language used by Japanese business people writing business letters at www.someya-net.com/concordancer/.

1. The terms *register* and *genre* are often interchangeable in corpus studies.

Nuances of Language

Another area that corpus linguistics addresses in language teaching is nuances of language; like ESP, nuances of language are also a sort of combination of the areas of language teaching addressed by corpus linguistics that have already been discussed. This is the contribution to teaching that, in my experiences, many language instructors seem to appreciate the most.

Nuances of language refer to questions that students might ask that we just don't know the answers to. Often, the questions specifically relate to areas of collocation and frequency. Our response, as teachers, is usually something like, "There is no difference/such thing" or "That's just the way it is in English," which isn't particularly helpful. For example, what if a student says, "I stubbed my large toe." That doesn't sound exactly right. Why? Well, that's just not the way we say it in English. When do you use *is not* (or *'s not*) versus *isn't*? We're likely to tell students that there's probably not really any difference between the two. But, corpus linguistics can answer both these questions. McCarthy (2004, p. 4) found that *is not* is used more with pronouns and *isn't* with nouns. And large toe? We use *large* to describe quantity and *big* to describe physical size, so *big toe* it is.

Syllabus Design

The final area of language teaching that corpus linguistics addresses is syllabus design. Phraseology, lexicogrammar, register, ESP, nuances of language—all of these areas can be used to more accurately and effectively design syllabi by helping us see what students really need to know about language: frequency and collocation for vocabulary, grammar patterns for different registers, and specific knowledge for specific purposes. And what are accurate descriptions of it. For example, the present perfect appears in almost every grammar textbook. It's usually defined as "recent past" or "completed action." However, a corpus study revealed that more than 80 percent of the time, present perfect is used to signify "indefinite past" (Mindt, 2000, p. 224). Statistics of this kind help textbook writers, course designers, and teachers set priorities for the classroom. If you, as a teacher, are armed with this kind of knowledge, you can supplement course materials with information that is relevant for students.

TOOLS

Types of Corpora

A corpus is a principled collection of authentic texts stored electronically that can be used to discover information about language that may not have been noticed through intuition alone. When you want to consult a corpus, what exactly should you look for? This is a very important question.

Because most published materials based on corpora make use of large, general corpora, many readers may believe this is the type of corpus that can be useful in the classroom. Actually, there are approximately eight types of corpora—generalized, specialized, learner, pedagogic, historical, parallel, comparable, and monitor—and which type should be used depends on the purpose for the corpus; only the four types of corpora that are most useful for employing the corpus approach directly in the classroom will be discussed here.

Generalized Corpora

The broadest type of corpus is a generalized corpus. Generalized corpora are often very large, more than 10 million words, and contain a variety of language so that findings from it may be somewhat generalized. Although no corpus will ever represent all possible language, generalized corpora seek to give users as much of a whole picture of a language as possible. The British National Corpus (BNC) and the American National Corpus (ANC) are examples of large, generalized corpora. The COCA is also an example of a generalized corpus. These large, generalized corpora contain written texts such as newspaper and magazine articles, works of fiction and nonfiction, as well as writing from scholarly journals; these corpora also contain spoken transcripts such as informal conversations, government proceedings, and business meetings. If generalizations about language as a whole are to be drawn, a large, general corpus should be consulted.

Specialized Corpora

A specialized corpus contains texts of a certain type and aims to be representative of the language of this type. Specialized corpora can be large or small and are often created to answer very specific questions. Examples of specialized corpora include the Michigan Corpus of Academic Spoken English (MICASE), which contains only spoken language from a university setting; the CHILDES Corpus (MacWhinney, 1992), which contains language used by children; the MICUSP, Michigan Corpus of Upper-level Student Papers, a collection of papers from a range of university disciplines; and a medical corpus containing language used by nurses and hospital staff. Specialized corpora are often used in ESP settings. The AWL, for example, was generated from a specialized corpora of academic texts.

Learner Corpora

A learner corpus is a kind of specialized corpus that contains written texts and/or spoken transcripts of language used by students who are currently acquiring the language. Learner corpora are often tagged and can be examined, for example, to see common errors students made. A well-known learner corpus is the International Corpus of Learner English (ICLE) (Granger, 2003), which contains essays written by English language learners with 14 different native languages. While the ICLE is more generalized, containing writings from learners with 14 different native languages, other learner corpora are more specialized; for example, the Standard Speaking Test Corpus (SST), comprised of oral interview tests of Japanese learners. Targeted instruction can be developed for general language teaching or for specific language groups depending on the type of learner corpus. Chapter 7 will look at corpus-designed activities created from a learner corpus.

When a corpus is **tagged,** each word included in the corpus has a marker added to it that gives additional information. Often, tags are part of speech markers, enabling users of corpora to search not only for specific words, but also for specific words used as a particular part of speech.

Pedagogic Corpora

A pedagogic corpus is a corpus that contains language used in classroom settings. Pedagogic corpora can include academic textbooks, transcripts of classroom interactions, or any other written text or spoken transcript that learners encounter in an educational setting. Pedagogic corpora can be used to ensure students are learning useful language, to examine teacher-student dynamics, or as a self-reflective tool for teacher development.

Creating Corpora

A corpus is a principled collection of authentic texts stored electronically. When creating a corpus, there must be a focus on three factors: the corpus must be principled, it must use authentic texts, and it must have the ability to be stored electronically.

A corpus is principled, meaning that the language comprising the corpus cannot be random but chosen according to specific characteristics. Having a principled corpus is especially important for more narrow investigations; for example, if you want your students to look at the use of signal words in academic speech, then it's important that the corpus used is comprised of only academic speech. A principled corpus is also necessary for larger, more general corpora, especially in instances where users may want to make generalizations based on their findings. In creating the Longman Spoken and Written English Corpus (LSWEC), a 40 million–word corpus created to identify and understand grammatical patterns in English—the corpus that the information in the *Longman Grammar of Spoken and Written English* (Biber et al., 1999) is based on—

the compilers included a representative sampling from conversation, fiction, news, and academic prose. Whatever the purpose of the corpus, it must be principled.

A corpus must also include authentic texts. Although there is debate over the definition of "authentic" texts in second language teaching (see Widdowson, 1990, for example), for purposes of this discussion, authentic texts are defined as those that are used for a genuine communicative purpose. In the MICASE, only speech acts that naturally occurred in the course of routine daily events at a university are included. The LSWEC includes texts from daily newspapers that were distributed and conversations that took place during participants' weekly routines. The main idea behind the authenticity of the corpus is that the language it contains is not made up for the sole purpose of creating the corpus.

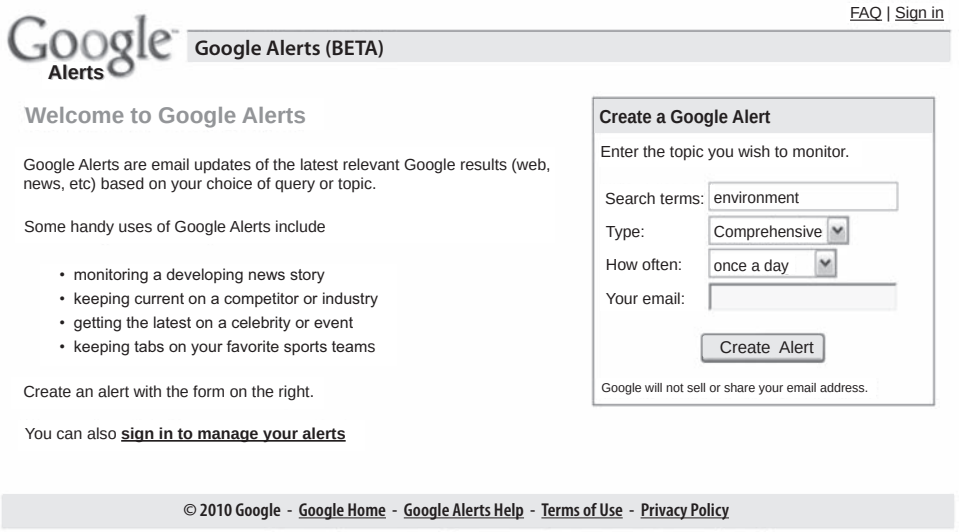
Lastly, a corpus is stored electronically. Corpora can be saved in text format (.txt), rich text format (.rtf), and/or web-based format (.html), or others, depending on the concordancing program used to access texts. The electronic storage and easy accessibility of texts is one of the major factors that allows corpus linguistics to be applied in the classroom.

If you are creating your own corpus, one way to gather principled, authentic texts that can be stored electronically is through Internet "alerts." Search engines such as Yahoo and Google gather email updates of the latest relevant results based on a topic or specific query generated by the user. Alerts can be used to monitor a developing news story, keep current on a particular theme, get the latest on a celebrity or event, and/or keep tabs on a favorite sports team; an example is shown in Figure 1.1. One way to use corpora created from alerts is to investigate common vocabulary used in certain topics, such as frequent content words used in articles that discuss the environment.

Another means of gathering principled, authentic texts that can be stored electronically is looking at Internet essay sites. Many of the academic essay sites have a disclaimer that their essays should be used for research purposes only, and should not to be downloaded or turned in as one's own work. These sites can be very helpful for creating corpora specific for academic writing with term papers, essays, and reports on subjects such as business, literature, art, history, and science. You can even access essay sites that aren't academic. The creation of a corpus using essays from NPR's "This I Believe" program can be analyzed for American viewpoints and language, for example.

Corpora can also be created from resources at hand. Textbooks can be used to create a pedagogic corpus to investigate the language of academic textbooks. This would be especially useful for students enrolled in an Intensive English Program (IEP) or an English for Academic Purposes (EAP) program. Learner corpora can be created from the compilation of student work taken from one particular class, for one particular student, or from a series of students and classes. Students can analyze their own language use and pinpoint areas that need further instruction or document progress that has been made.

Figure 1.1
Google Alerts



Google Alerts allow users to collect email updates based on a specific topic or query as shown here. You can choose register, news, blogs, the web, videos, groups, or all of the above (as shown). Either once a day (as shown), as it happens, or once a week, text will be emailed that relates to a specified topic. These texts can be combined to investigate vocabulary or grammar patterns, for example, for particular themes.

An important aspect related to creating corpora is the issue of copyright, especially if findings from a corpus will be distributed via a handout or published in any form. Contact Internet sites for their permissions policy, and always get students' written permission before using their work for any purpose. Some institutions may also require you to complete an Internal Review Board (IRB) application for a corpus study.

An institution's IRB serves to monitor research involving human subjects. The role of the IRB is paramount in medical studies during which physical or psychological damage may be done to research participants, although studies involving research of normal educational practices—such as those concerning instructional strategies or the effectiveness or comparions of instructional methods—can often be exempted so long as a clear demonstration can be made that human subjects will not be identified (e.g., the use of numbers instead of names or the keeping of research documents in locked drawers). **Copyright is a serious issue and should not be overlooked.**

Concordancing Programs

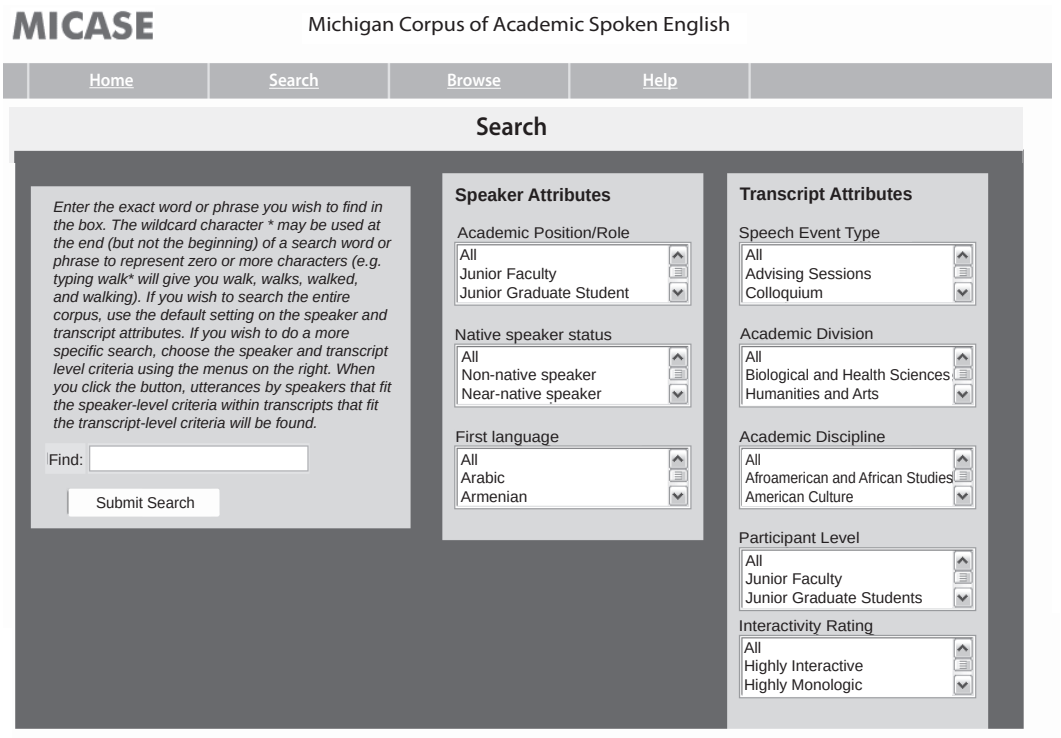
Two tools are needed to effectively apply corpus linguistics in the classroom: the corpus and the concordancing program. Concordancing programs are computer software used to access and sort data from the corpus. Most large corpora, like the MICASE, have a built-in concordancing program. When the concordancing program is built in,

as is the instance with the MICASE, only the search options are important to understand for the program. See Figure 1.2 for search options in the MICASE concordancer.

When you create your own corpora or want to access a corpus that does not have a built-in concordancing program, it's important to note that many effective concordancing programs are available. Some concordancing programs are very affordable, and others are free.

A concordancing program should be chosen based on what information you want from your corpus. On the basic level, all concordancing programs provide a frequency list that provides each word used in the corpus as well as the number of times the word appears in the corpus. Often, the list can be sorted by frequency or alphabetically. In addition, most concordancing programs show concordance lines from the corpus. Concordance lines are all the instances of a word or phrase in the corpus. Many concordancing programs will allow you to sort the concordance lines according to what comes before or after a specified word or phrase and will let you see more extended context of the word as it appears in the corpus.

Figure 1.2
The MICASE Corpus



MICASE allows users to search for a word or phrase within the corpus. You can also narrow your search by specifying various speaker attributes or transcript attributes, so that the concordancing program will only display results of the word or phrase you are searching that are found within those attributes.

Source: http://micase.elicorpora.info. Used with permission.

An example of this type of basic concordancing program is TextSTAT 2.8 (Hüning, 2008). TextSTAT, Simple Text Analysis Tool, can be downloaded free from the Internet and has all the basic features needed to access a corpus: uploading of files to create a corpus, retrieving of word forms, viewing of concordance lines, and accessing of extended context. These basic features allow users to retrieve word frequencies from the corpus and sort alphabetically (or by frequency or by retrograde, alphabetically backward), to establish a minimum/maximum frequency, and to search for words containing affixes, as well as search and view concordance lines and extended context. WordSmith Tools (Scott, 2004) and MonoConc Pro (Barlow, 2007) are other concordancing programs that can be downloaded from the Internet. WordSmith Tools and MonoConc Pro do have a cost, but a personal license (for use on your personal computer) is affordable.

Download **TextSTAT** at http://neon.niederlandistik.fu-berlin.de/textstat/.

PROCEDURES

A Framework for Creating Corpus-Designed Activities

Parts 2 and 3 of this book focus on the applications of corpus linguistics to language teaching, with many of the chapters (Chapters 4–7) focusing specifically on corpus-designed activities that can be used in your classroom.

As shown in Table 1.2, creating corpus-designed activities involves seven steps.

Table 1.2
A Framework for Creating Corpus-Designed Activities

▪ Ask a research question.
▪ Determine the register on which your students are focused.
▪ Select a corpus appropriate for the register (or compile authentic texts from that register).
▪ Utilize a concordancing program for quantitative analysis.
▪ Engage in qualitative analysis.
▪ Create exercises for students.
▪ Engage students in a whole-language activity.

Ask a research question.

The "research question" for a corpus-designed activity could resemble, "What's the difference between *through* and *between*?" which was discussed earlier in this chapter. Or, "How do you use signal words in academic speaking?" which will be explored in Chapter 5.

Determine the register on which your students are focused.

As previously discussed, language is used differently in different registers. It's important when creating a corpus-designed activity that you know which register is relevant for your students. If students are practicing informal conversation, looking at a corpus of academic papers won't be helpful.

Select a corpus appropriate for the register (or compile authentic texts from that register).

Whether you create your own corpus or you use a corpus that already exists is not particularly relevant to creating a corpus-designed activity beyond the appropriate register and size of the corpus. What is important is that the corpus contains authentic language used for real-life communication.

Utilize a concordancing program for quantitative analysis.

Utilizing a concordancing program is the third step in the corpus approach. To create a corpus-designed activity, a concordancing program must be used to access the language stored in the corpus. If you are using a corpus that has a built-in concordancing program, be sure to understand all the search functions. If you are employing an outside concordancing program with a corpus, be sure that the program performs the functions you need to answer your research question.

Engage in qualitative analysis.

Qualitative analysis is the last step of the corpus approach. Most often, qualitative analysis will answer the question *Why*? For qualitative analysis, take the quantitative information given by the corpus's concordancing program and determine its significance.

Create exercises for students.

Preparing concordance lines and traditional fill-in-the-blank and gap-fill activities for students to examine and engage in is a central element of corpus-designed activities.

Engage students in a whole-language activity.

The grammar methodology of form, meaning, and use is also applicable to applying corpus linguistics in the classroom. The first and second steps of this methodology, form and meaning, are addressed through corpus-designed activities; but when applying corpus linguistics in the classroom, you should also ensure that students have an opportunity to use the target feature under investigation. Creating gap-fill exercises and whole-language activities will provide such opportunities for students and encourage acquisition of the target features at hand. Examples of these types of activities are included in Part 3.

Look again at the steps for creating corpus-designed activities listed in Table 1.2. What do you notice? The steps of the process are not numbered; this is because they may not take place specifically in this order. You may start with a research question, a corpus, a register, or even a whole language activity.

As you read Chapters 4–7, refer back to this framework for creating corpus-designed activities; it will help you make the connection from the reading and exercises to creating your own activities.

Modifying Activities by Language Level

Most of the activities discussed in Part 3 require at least an intermediate to high-intermediate level of English, but there are ways to modify corpus-designed activities to make them more accessible to your students. By their nature, corpus activities are for more advanced levels, but they can be adapted for students at lower levels.

To modify corpus-designed activities for low (beginning to low-intermediate) language level:

- Ask simple research questions.
- Find your own concordance lines.
- Adapt lines for students' level.
- Present students with fewer lines.
- Encourage group or whole class work.

Ask simple research questions.

General questions that have to do more with language as a whole are usually less ambiguous and more significant for lower-level students than specific questions like, "What's the difference between *through* and *between*?" For example, if you are teaching a group of Hispanic students who consistently use *have* to tell how old they are, a simple activity with *have* versus *am* when giving age can be very enlightening for learners.

Find your own concordance lines.

Instead of utilizing a whole corpus of authentic English for lower-level students to navigate, look for instances of the pattern you would like to investigate in materials around you, like the textbook, a newspaper, or magazine.

Adapt lines for students' level.

Materials that you intend to use in your corpus-designed activity can be modified for sentence structure and vocabulary so long as the feature under investigation remains intact. For example, these following sentences are used in the Chapter 5 materials on using *though* in academic speaking and exemplify the notion that *though* can be used to show a contrast in ideas.

> The two disciplines do not appear on the surface to have very much in common. Historically, though, anthropologists and epidemiologists have worked together for a very long period of time.

The sentence structure and vocabulary in these two sentences is quite advanced for low-intermediate learners, but they can be modified and still demonstrate *though* showing a contrast in ideas:

> The study of culture and the study of people's health do not seem similar, though they have been studied together for a very long time.

While some corpus linguists may argue against this method, I believe it's a realistic way for lower-level students to take advantage of the benefits of learning from corpus-designed activities and outweighs any drawbacks of working with adapted language.

Ask students to work with fewer lines.

Students don't necessarily have to study 20–50 lines in a corpus-designed activity; 10 lines may be enough for lower-level students to study, especially in investigations dealing with more general ideas, such as *have* versus *am,* as previously discussed.

Encourage group or whole class work.

Completing a corpus-designed activity as a whole class is another great way to give lower-level students exposure to the benefits of learning from corpus-designed activities without being too challenging for their language level.

ON YOUR OWN

For many instructors who want to provide the best possible instruction for their students but are pressed for time and resources, theoretical principles are often a luxury. What makes corpus linguistics work, so to speak, are the practical activities that can be used with students in an everyday classroom. Knowledge of what corpus linguistics is and is not, questions that corpora can answer, the corpus approach, types of corpora and concordancing programs, and how to create corpus-designed activities all help to provide a solid foundation for understanding the applications of corpus linguistics. Try these activities to get an idea of how you can apply corpora to your language learning classroom.

1. Go to www.collins.co.uk/corpus/CorpusSearch.aspx to view the top 100 collocations of a word or words.
2. Visit http://corpus.byu.edu/bnc/x.asp to view a side-by-side comparison of frequency and collocation in spoken versus written language.