12

CORPUS-BASED VOCABULARY SUPPORT FOR UNIVERSITY READING AND WRITING

Mark Davies and Dee Gardner

Vignette

Imagine that you are Susana Meléndez, a nonnative speaker from Peru, who is enrolled in an advanced writing course at a university in the United States. As you write a paper for your class, you have the following questions: Which is more common—potent argument or powerful argument? How about utter despair or sheer despair? What verb should I use with havoc—make, wreak, or produce? Which sounds better in academic English—brainy or knowledgeable? Is (urban) sprawl a good thing or a bad thing? Does scheme have a different meaning in American and British English? Do I use excel in or excel at? Would informal writing use should talk or must talk in American English (and what about British English)? Is it have strived or have striven in American English? Assuming your dictionary and thesaurus can't answer these questions, where can you go to get quick answers as you write your paper?

Introduction and Overview of the Challenges

Academic vocabulary knowledge has been identified as a key component of academic literacy skills (Biemiller, 1999; Corson, 1997), which, in turn, have been strongly correlated with academic success, economic opportunity, and societal wellbeing (Goldenberg, 2008; Jacobs, 2008). It is also widely recognized in English-speaking countries like Great Britain, the United States, Australia, Canada, and New Zealand that higher education students from non-English-speaking backgrounds face a formidable challenge in trying to acquire adequate levels of English academic literacy (Henry & Roseberry, 2007). The complex demands placed on such learners has sparked a number of important technological innovations, one of these being the use of electronic corpora to support academic literacy needs.

The advent of high-powered technology and searchable electronic corpora has recently inspired a host of data-driven resources for teaching and learning English. Examples of the range of corpus applications include (1) large megacorpora such as the Corpus of Contemporary American English (COCA) and the British National Corpus (BNC); (2) corpora dealing with language change over time (e.g., the Corpus of Historical American English [COHA]—Davies, 2012); (3) corpora based on actual learner language, or Learner Corpora (e.g., Gilquin, Granger, & Paquot, 2007); (4) corpora used to determine core academic words and phrases (e.g., Gardner & Davies, 2014; Simpson-Vlach & Ellis, 2010); and (5) specialized corpora such as those used to investigate the language of engineering (e.g., Mudraya, 2006), agriculture (Martínez, Beck, & Panza, 2009), biochemistry (Kanoksilapatham, 2005), business (e.g., Blanpain, Heyvaert, & Laffut, 2008; Nelson, 2006), history (Cortes, 2004), medicine (Wang, Liang, & Ge, 2008), law (Hafner & Candlin, 2007), and many other areas of academic literacy. All of these corpus applications attest to the fact that data-driven methodologies have found a home in contexts where academic English is being studied.

The appeal of corpora for language training is that they represent the way people really write or talk, rather than the textbook examples we often find in traditional course materials, or explanations about language usage based on our intuitions, which are often inaccurate (Hunston, 2002). While corpus-based vocabulary resources such as word lists have been with us for nearly 80 years—thanks to tedious manual calculations and rudimentary computer analyses in the early days of linguistic computing—the recent introduction of machine-searchable corpora and the availability of powerful personal computers, smartphones, and other electronic devices has also made it possible to bring data-based applications directly to classrooms and individual learners (Aijmer, 2009; Bennett, 2010; Gardner, 2012; Granger, Hung, & Petch-Tyson, 2002; Reppen, 2010).

However, Bernardini (2004) makes an important distinction between "uses of corpora as sources of descriptive insights relevant to language teaching/learning and uses of corpora that directly affect the learning and teaching process(es)" (p. 15). While the line between these two uses of corpora may be blurred at times, it is perhaps helpful to think of the distinction as "resources" versus "methods." The purpose of this chapter is to show how COCA and its online interface, WordAndPhrase.info, can be used as a "resource" to support the academic literacy needs of English language learners in higher education, particularly in terms of vocabulary usage (both words and phrases), understanding of genre (Flowerdew, 2005), and knowledge of grammatical/syntactical patterns (Hunston, 2002). Actual corpus-based classroom methods that have been shown to directly affect learning and teaching will not be discussed, as such methods are in their relative infancy and lack an adequate body of empirical support (Aijmer, 2009; Bennett, 2010). However, it is hoped that our discussions and

examples in this chapter will encourage advanced language learners and their teachers to begin to see the possibilities for learning and teaching that corpus searches can provide.

Implications and Applications

Consider again the many different challenges faced by our hypothetical university student, as posed in the vignette at the beginning of this chapter. Where could she go to get answers to these questions—a dictionary, or perhaps a thesaurus? Actually, very few of these questions could be answered with either type of resource. What she needs is the ability to quickly search millions of words of text from different styles and genres of English to determine what sounds the most natural.

A corpus can provide this type of insight. Since the 1980s, there have been available a number of corpora—such as the BNC and the Bank of English—which allow language learners to see what is "really happening" in the language. Unfortunately, neither of these two corpora is currently being updated, and neither corpus focuses on American English. Since 2008, however, the 450 million-word COCA (see Davies 2009, 2011) has been available, and it can help advanced ESL students and their teachers to answer the kinds of questions most English language learners face.

In addition to the regular COCA interface (corpus.byu.edu/coca), a number of other COCA-based resources—also very relevant for language learners—have also become available (see corpus.byu.edu). Perhaps the most useful of these is www.WordAndPhrase.info, which will also be discussed in this chapter. In the following sections, we will consider how all of these corpus-based resources can be used to answer many different types of questions that nonnative speakers of English and their teachers might have, as they deal with the literacy demands of academic English in higher education.

Using Collocates to Find the Meaning of Words

COCA allows language learners to quickly find the collocates (nearby words) of a given word or phrase, and these collocates can provide very useful insight into the meaning and usage of a word. For example, what is the meaning of *break*, or "what do we break"? In order to find out, users simply input the word, (optionally) specify the part of speech for the collocates, and then click to find the collocates. In less than two seconds, users can see a list of words like *law* (1,527 tokens near *break*), *heart* (1,454), *news* (1357), *record* (995), *rules* (943), *silence* (896), *ground* (804), *leg* (567), *barriers* (486), *cycle* (468), and *pieces* (445). Users can also choose to find collocates that occur much more frequently with *break* than their overall frequency in the corpus might suggest, which often indicates that the two words (*break* + collocate) have an idiomatic sense. For example, when users sort by

"relevance," they find that the top collocates are *logjam* (83 occurrences), *deadlock* (127), *monotony* (71), *stranglehold* (48), *taboos* (46), *impasse* (75), *stalemate* (66), and *barrier* (398), most of which have a strongly idiomatic feel to them. (See Davies and Gardner 2010 for the most frequent collocates of the top 5,000 words in English.) Of course, obtaining such rich output so quickly does not guarantee that students will know how to quickly sort through the data to draw conclusions, so they will need plenty of practice (both teacher-directed and independent) to become comfortable with interpreting the results of corpus searches. They should also be encouraged to utilize the tutorials provided in the COCA interface to understand how to actually perform various searches.

So how could nonnative speakers use this information, as they develop advanced proficiency in reading or writing? The answer is that in many cases, the collocates provide insight into meaning and usage that can't be found in even the best of dictionaries. We will provide two quick examples. First, consider the word *brooding*. A typical dictionary entry might indicate that the word means "cast in subdued light so as to convey a somewhat threatening atmosphere" (dictionary.com). On the other hand, consider the collocates of this word in COCA: (noun) *dark, eyes, look, silence, presence, sky, sense, cloud, thought, mood, portrait, bird*; (misc.) *dark, over, sit, silent, heavy, gray, stare, handsome, mysterious, beneath, moody*. Most would agree that the collocates "paint a picture" of the sense of this word that is far beyond what a dictionary can produce.

Consider a second example: the collocates (and thus the meaning and usage) of the word *sprawl*. The site www.dictionary.com indicates that (as a noun) this word means "the act or an instance of sprawling" or "a sprawling posture," neither of which is overly insightful. COCA, on the other hand, provides the following collocates: (adjective) *urban, suburban, rural, industrial, metropolitan, vast, unchecked, surrounding, Southern, increasing*; (noun) *city, development, traffic, growth, pollution, congestion, land, town, farmland, county*; (verb) *create, encourage, stop, fight, reduce, curb, slow, threaten, limit, crawl*. As we can see, the collocates show that *sprawl* refers particularly to the growth of cities (*city, suburban, farmland*), that it may be more common in the *Southern* United States, that it is associated with *pollution* and *congestion*, and that people are trying to *reduce, stop*, and *fight* against it. In summary, collocates "paint a picture" of a word that is far beyond what virtually any dictionary can provide.

Using Collocates to Compare Synonyms

One of the most useful aspects of collocates is their ability to show "shades" of differences in meaning between two near synonyms and therefore help language learners to use the most appropriate of the two synonyms. And this information, which can be easily obtained from a corpus, is not the type of detail that would be found in a typical thesaurus, which simply provides lists of words that—in principle—could be substituted for each other.

For example, consider the two near synonyms *small* and *little*. For a student who is a native speaker of Spanish, these two words look quite interchangeable, and relate to the Spanish word *pequeño*. It would therefore be quite difficult for such a student to know which of the two English words are used with different collocates. But with the COCA interface, in a matter of three to four seconds, users can begin to observe that the following collocates are used primarily with *little*: *quantities, percentage, populations, amounts, sizes, fraction* (all of which refer in some way to quantities and ratios), whereas the following collocates are used primarily with *small*: *while, league, luck, sleep, fun, attention, sympathy, sister, doubt, sympathy, help* (most of which refer to abstract nouns). So a language learner who is struggling to find the right word—*little* or *small*—to use with a given collocate could find, with some practice, which of the two is more frequent. It is also important to note here that simply asking the teacher this same question is likely to be unproductive, as such information generally is not explicitly known by teachers, and is often not intuitive.

Consider a second example. A nonnative speaker of English wants to know whether you *rob food* or *steal* it. Do you *steal* or *rob* someone's *identity*? And do you *steal* from or *rob* someone at *gunpoint*, or *rob* or *steal* from someone *blind*? With the corpus, these are easy questions to answer. Collocates that occur much more frequently with *steal* are *money, food, cars, wallet, identity*, and *drugs* (where the noun is the thing that is directly taken). The collocates of *rob*, on the other hand, tend to refer to the person or institution from which things are stolen (*bank, store, victim, person, restaurant, children*), or they refer to the circumstance or way in which the item was taken (*gunpoint, sleep, blind, rape, place*).

Finally, consider a case like the near synonyms of *complete*: *total, sheer*, and *utter*. Again, if students were to consult a thesaurus, they would simply see a list of words, with little, if any, indication of what the differences between them might be that might help them select the most appropriate word in a given context. But a quick search in COCA shows quite nicely that the collocates of *sheer* (but not *utter*) are *number(s), volume, force, size, weight, scale, magnitude* (referring to quantity) as well as *luck*, and also *cliff(s), rock, face*, and *drop* (referring to a steep drop-off). The collocates of *utter* are perhaps even more interesting: *darkness, failure, destruction, disregard, contempt, fool, desolation, defeat, silence, absence, disbelief, hopelessness, disaster*, and *defeat*. All of these are very negative, reflecting the strong negative "prosody" of this word. When learners see such a list, they are able to understand that to use *utter* is to use a word that is very heavily loaded with a negative sense, which is something that shows up very nicely in a corpus, but which would likely not show up at all had they consulted only dictionaries or thesauri.

Comparing Synonyms Directly

In a simplistic thesaurus-based approach to synonyms, learners simply see a long list of words, with no sense of which synonyms are actually used in the "real world,"

and particularly in different styles of speech and writing. For example, learners might see that *perambulate* and *saunter* are "synonyms" of *walk* (verb), and therefore they might be tempted to write or say things like "So, when did you perambulate to class today," or "When I was young, I often sauntered to class by myself."

The COCA web interface is unique among corpora in that it allows users to quickly find the frequency and use of the synonyms of a given word, and also see how and how much they are used in different genres, a technique that avoids the problems of a simple thesaurus-based approach. For example, suppose that students want to see the synonyms of *precarious*. They would simply enter [=precarious] in the search form, and they would then see something like the results presented in Table 12.1, which shows the overall frequency of each synonym and its frequency in each genre. The fact that *precarious* is used only about one-tenth as much as *weak* or about one-sixth as much as *slight* could suggest to the learner that this word has a much narrower range of meanings and uses. (Note that on the corpus website there are 17 synonyms—only 10 are shown here—and the cells are colored to show relative frequency, whereas this is more difficult to see in the grayscale Table 12.1.)

One of the nice features of the interface is that students can "explore" a "chain" of meaning by simply clicking on the [S] after any synonym to see the synonyms of that word, and then clicking on another synonym in the new set, and so on. For example, if students click on *precarious*, they would then see *dangerous, uncertain, risky, hazardous, unstable, shaky*, and *unsafe* (among others), and they could then click on *shaky* to see *uncertain, trembling, questionable, unstable, dubious, doubtful, unreliable*, and so on. In this way, the students can follow through the chain of meaning, from one sense to another. And in each case, they could see the frequency of the synonyms and their distribution by genre to

TABLE 12.1 Synonyms of *Precarious* by Genre

Synonym	*Total*	*SPOK*	*FIC*	*MAG*	*NEWS*	*ACAD*
Weak [S]	16,176	2,190	3,816	3,266	2,820	4,084
Slight [S]	9,169	796	3,576	2,040	1,293	1,464
Delicate [S]	8,377	714	2,932	2,333	1,476	922
Fragile [S]	5,916	738	1,549	1,496	1,060	1,073
Unstable [S]	3,097	458	339	670	390	1,240
Shaky [S]	2,493	323	856	523	609	182
Frail [S]	1,986	188	898	347	244	309
Brittle [S]	1,679	79	730	491	234	145
Precarious [S]	1,612	182	299	334	280	517
Tenuous [S]	1,364	126	219	266	255	498

know whether a word has a more general and frequent use (or whether it is more specialized), and whether it is more informal or more formal.

The synonym feature is the most useful for learners when it is used in a particular context, like a writing assignment. For example, suppose that students are considering using the phrase *potent argument* in their papers but want to see whether there are better, more common ways to express this. The students would simply enter [=*potent*] *argument* into the search form, and they would then see *strong argument* (138 tokens), *powerful argument* (81), *convincing argument* (81), *persuasive argument* (63), *effective argument* (18), *vigorous argument* (7), *potent argument* (6), *influential argument* (5), and *forceful argument* (5)—all of which would probably suggest that there are better alternatives to *potent argument*. Of course, any number of examples like this could be given. The point is that the corpus can pinpoint just the right synonym to combine with a given word (and in a given genre), which is something that even the best thesauri cannot do.

The Importance of Genre

As we have seen, the appropriateness of a word is very much dependent on the genre in which it is used. If students are writing for an academic audience, then certainly they need to keep that in mind as they select from among the possible words and phrases they could use. Unfortunately, this type of genre-based information is typically not available in a dictionary, but, fortunately, it is quickly available via a corpus.

As a simple demonstration of this, in COCA it is possible to find all words that are more common in one genre than in another. For example, one simple search shows that the following verbs occur much more in fiction than in academic writing—*whimper, snore, waltz, squeal, slump, twirl, perk, snuggle, quiver, sob, claw, saddle,* and *croak*—while verbs that are much more common in academic than fiction writing include *operationalize, remediate, predominate, aggregate, reformulate, reconceptualize, abrogate, individualize, facilitate, marginalize,* and *reify*. And of course there are important differences in genre at the phrasal level as well. For example, a simple search shows that the following phrasal verbs are much more common in academic than in fiction writing: *bear out, contract out, phase out, emerge out, opt out, carry out, rule out, map out, separate out,* and *sketch out*. The search additionally shows that the following phrasal verbs are much more common in fiction: *stare out, freak out, scream out, pop out, make out, bust out, chill out,* and *chicken out*.

Learners can of course limit their search to just a specific set of words, such as synonyms of a given word. Suppose, for example, students want to see which synonyms of *strong* are used in different genres. With one simple search, they could see which synonyms of *strong* are much more common in an informal genre like fiction (e.g., *beefy, burly, strapping, spicy, pungent, brawny, well-built, biting, sturdy, dazzling*) and which are much more common in academic writing (e.g., *effective, deep-seated, clear-cut, durable, compelling, robust, persuasive, dedicated,*

potent, powerful). This would hopefully serve as a clear reminder that students would not use the phrases *burly argument* or *spicy support* in an academic paper, or expect to see *deep-seated hands* or *compelling wind* in a short story.

Of course, many times learners simply want to know "Is this word an academic word, or not?" Such judgments are readily available to native speakers, but they are typically not for language learners. Therefore, it is very useful to simply input a given word or phrase and see how frequent that word or phrase is in different genres. For example, consider the frequency charts for *sustain* and *withstand* in Figure 12.1, which show that *sustain* is a much more academic word, whereas *withstand* is more evenly distributed across genres. (Note: the first row indicates the number of tokens in each genre per million words, and the second row shows the normalized frequency in each genre per million words.)

Learners can see the genre frequency of any single word or phrase or compare any two contrasting words (as in Figure 12.1). This can extend even to quasi-grammatical lexical choices, as in Figure 12.2, where we see that *have to* is relatively informal (e.g., *they have to leave*), *should* is more evenly distributed across all levels of formality (e.g., *they should leave*), and *must* is found primarily in formal, academic writing (e.g., *they must leave*). Again, native speakers can often "intuitively" sense such differences, but the corpus data can be invaluable to nonnative speakers to help them know which words are (in)formal and which ones are not.

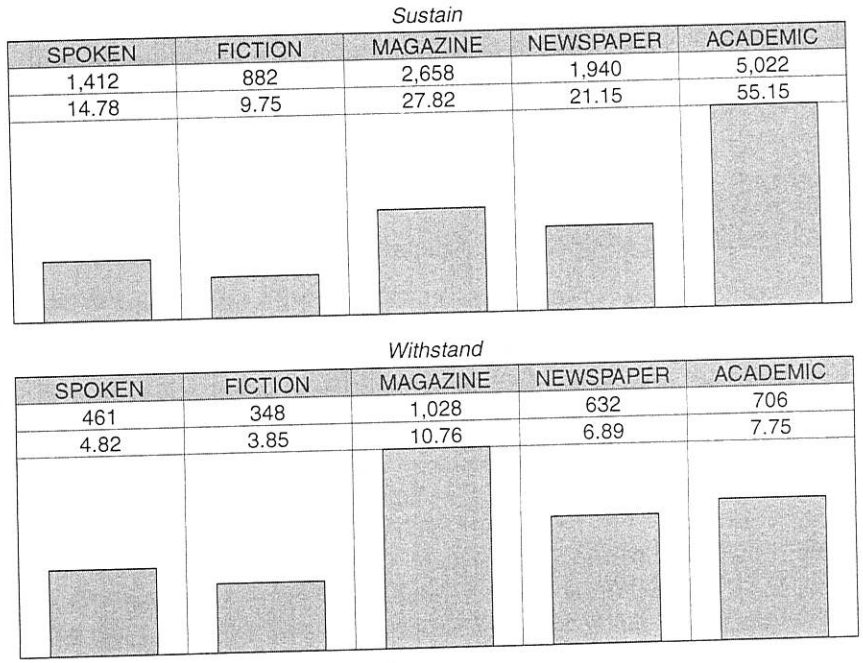

FIGURE 12.1 Frequency of *Sustain* and *Withstand* in COCA, by Genre.

Have to

SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
96,221	47,166	40,008	41,096	15,386
1,006.86	521.58	418.67	448.07	168.95

Should

SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
88,828	55,890	67,961	59,607	81,919
929.50	618.05	711.20	649.90	899.55

Must

SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
21,909	44,586	35,694	28,865	59,801
229.26	493.05	373.53	314.72	656.68

FIGURE 12.2 Frequency of *Have to, Should*, and *Must* in COCA, by Genre.

Genre-based information can be particularly useful for language learners as they consider idioms, many of which are of course more common in the more informal genres, and are typically not as acceptable in formal, academic genres. Dictionaries typically do not indicate the highly informal nature of these idioms, but the charts from a balanced corpus can indicate this quite well. Consider, for example, the following three idioms with the word *head*: *use one's head* (e.g., *he told me to be careful and use my head*), *in over one's head* (e.g., *he was way in over his head on that project*), and *off the top of one's head* (e.g., *but off the top of my head, I can't imagine why*). In all cases, the idiom is much less frequent in formal or academic English, as can be seen in Figure 12.3.

Use one's head

SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
37	71	57	34	3
0.39	0.79	0.60	0.37	0.03

In over one's head

SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
37	71	57	34	3
0.39	0.79	0.60	0.37	0.03

Off the top of one's head

SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
90	46	21	25	6
0.94	0.51	0.22	0.27	0.07

FIGURE 12.3 Frequency of Several Idioms with *Head* in COCA, by Genre.

Notice also the importance of corpus size with idioms. If we just had a small 4–5 million word corpus, we could only have about one-hundredth of the number of tokens that we have in COCA (which is currently 450 million words in size), and so for each of these idioms we would have just two to three tokens—certainly not enough data to help learners.

COCA and its associated suite of corpus tools at corpus.byu.edu can also be used to investigate dialectal differences at the macro-register level—such as differences in vocabulary usage between British (BNC) and American (COCA) dialects of English—as well as historical issues, such as determining words and phrases that sound overly "new" or "old-fashioned." Space constraints in this

chapter do not allow us to pursue these topics any further, but we encourage learners and their teachers to explore the possibilities of informing literacy practices that these resources provide.

Morphological and Syntactic Issues Related to the Lexicon

All of the discussion to this point has focused on the meaning and frequency of words and phrases. Words, of course, also have morphological and syntactic characteristics, and here again corpora can provide information that might not be readily available elsewhere.

Consider first the morphological characteristics of a word. For example, once students have determined that they want to use *strive*, should they use *have strived* or *have striven*? An online grammar guide may list both forms, but only a corpus can show the students what is really happening in the language, where we see that (perhaps surprisingly) *striven* is still used in the majority of the cases (about 59% of the total), with the highest use in academic texts (see Table 12.2). We also see, however (and probably not surprisingly), that the regular form *strived* is increasing over time—to nearly three times in the period 2010–2012 (labeled [10–12]) what it was just in the early 1990s (90–94).

To take just one more morphological phenomenon that nonnative speakers might face, we know that comparative adjectives can take either the *–er* suffix (e.g., *cleaner, faster*) or use *more* (e.g., *more wonderful, more considerate*). Unlike the adjectives just listed, however, many adjectives are divided between the two forms—for example, *purer/more pure*, or *unhappier/more unhappy*. Again, a corpus can easily show us what native speakers prefer. For example, the following forms prefer *–er*: *prettier* (98%), *simpler* (96%), *clearer* (90%), and *purer* (85%), while the following forms prefer *more* + adjective: *more unhappy* (65%), *more tender* (92%), *more sincere* (98%), and *more likely* (99%).

There are also phraseological and syntactic properties associated with words. To take just one example—which is more common—*different than, different to*, or *different from*? In this case, it depends on dialect, genre, and the time period. In terms of dialect, *different to* is definitely much more common in British than in American English. In terms of change over time, *different than* is clearly increasing in American English, but is still quite rare in British English. And in

TABLE 12.2 Word Forms: Have + Strived/Striven

	Total	*SPOK*	*FIC*	*MAG*	*NEWS*	*ACAD*	*90–94*	*95–99*	*00–04*	*05–09*	*10–12*
Striven	70	3	11	9	13	34	25	15	12	14	4
Strived	49	6	5	11	12	15	9	11	7	11	11
% Strived	0.41	0.67	0.31	0.55	0.48	0.31	0.26	0.42	0.37	0.44	0.73

terms of genre, *different from* is preferred to *different than* (the newer form) in (American) academic texts.

Of course, it is unreasonable to expect language learners to master the nuances of such variation, but even a quick corpus search would point out, for example, that *different from* is still the most common form in American English, and it would also point out the contrast between *different than* and *different to* in American and British English. And the fact is—based on the server log files of queries done with COCA and the other corpora—that thousands of people each week *do* use COCA primarily to look at fine-grained phraseological issues (such as which preposition to use with a particular verb or adjective), since this information is often not found in dictionaries or other online sources.

A New Learner-Friendly Interface: www.WordAndPhrase.info

As useful as the corpora are for the types of searches that we have described above, there are two problems in terms of how learners might use these resources. First, assuming that learners have a 500-to-600-word paper they have written, they would need to copy and paste many individual "snippets" from the paper (e.g., 1-to-5-word strings)—one after another, all of which is quite time consuming. Second, because there is a fair amount of "power under the hood" in terms of what the corpus can do, there is a learning curve in using the COCA corpus interface (even though there are many context-sensitive help files, with sample queries).

In order to make things easier for language learners, we recently created a new site, www.WordAndPhrase.info, that is based on COCA data, but which has a much more simplified interface. Most important, it allows users to enter and analyze entire texts, rather than requiring them to enter many individual words and phrases, as with the regular COCA interface. We will discuss the ability for users to enter entire texts below. First, however, we will briefly examine how the new WordAndPhrase interface provides information on individual words.

Via the WordAndPhrase interface, users simply enter the word that they are interested in, and they then see a wide range of useful information for that word (see Figure 12.4). This information includes (1) synonyms of the word, any of which can be clicked on to see the entry for the related words; (2) definitions of the word; (3) a chart showing the relative frequency of the word in each of the nine academic subgenres in COCA; (4) the top collocates of the word, which provide useful insights into meaning, usage, and phrasal possibilities; and (5) up to 200 sample concordance lines from COCA, which can be re-sorted to see the patterns in which the word occurs. In other words, rather than having to do separate searches for synonyms, and then collocates, and then concordance lines, and then frequency information (including by genre)—as with the regular COCA interface—all of this information is provided at one time at

WordAndPhrase. Basically anything that COCA can tell users about a word is all displayed together, with extensive links from one word to another.

The second key feature of the site allows users to input a text of their choosing (perhaps a reading selection, or a paper that they have written) and then see frequency information for each word in that text. For example, in the sample passage in Figure 12.5 (a random news release from Science Daily, www.sciencedaily.com/releases/2012/09/120927144234.htm), users would see statistics showing that 60% of the words are in the top 500 words (lemmas) in COCA, with another 20% from words 501–3,000 in the core. A final 20% of the words are not in the top 3,000 lemmas in COCA. The interface also shows that about 8% of the words are "academic" words, meaning that the word occurs with at least twice the expected frequency (per million words) in COCA academic texts. (Note that in the printed grayscale version in Figure 12.5, it is difficult to distinguish core academic from subgenre-specific words, but the web interface allows us to use color-coding to make these distinctions much more transparent.)

At the most basic level, users can click on the different frequency bands (see the top of the figure: lemmas 1–500, lemmas 501–3,000, lemmas above 3,000, and also "academic"), to see the words from those lists. For example, by clicking on [>3,000], students would see the following words shown in Table 12.3, which provide a fairly nice summary of what the article is about.

Other sites, such as Tom Cobb's Compleat Lexical Tutor (www.lextutor.ca) offer similar functionality. But WordAndPhrase offers something quite useful, which is (to our knowledge) not available at any other site. This is the ability

SEE LISTS	FREQ RANGE	1-500	501-3000	>3000	ACAD	SAVE TEXT	HELP
	613 words	60%	20%	20%	8%		

The deserts of Utah and Nevada have not always been dry. Between 14,000 and 20,000 years ago, when large ice caps covered Canada during the last glacial cooling, valleys throughout the desert southwest filled with water to become large lakes, scientists have long surmised. At their maximum size, the desert lakes covered about a quarter of both Nevada and Utah. Now a team led by a Texas A&M University researcher has found a new water cycle connection between the U.S. southwest and the tropics, and understanding the processes that have brought precipitation to the western U.S. will help scientists better understand how the water cycle might be perturbed in the future.

FIGURE 12.5 Sample Display from WordAndPhrase, Showing Frequency Information.

TABLE 12.3 Keywords from Inputted Text

[5 tokens] glacial; [4 tokens] southwest, wet; [3 tokens] cycle, desert, intervals; [2 tokens] caps, inland, precipitation, sediments, shorelines, tropical, tropics; [1 token] altered, archaeological, archived, assembling, buried, coastal, cores, cycled, deer, deserts, enhancing, evaluated, geological, geologists latitude, marine, migration, monsoon, mystery, oceanography, paleoclimate, perturbed, pollen, profoundly, progression, radiocarbon, shelters, southeast, speculation, strengthen, surmised, synthesized, weak, wildfowl	

FIGURE 12.4 Sample Display from WordAndPhrase for *Coherent*.

to click on any of the words in either of the customized lists (e.g., Table 12.3) or any of the words in the original text (e.g., Figure 12.5), and then see the full-featured entry for that word (e.g., Figure 12.4), including synonyms, definition, frequency information (including by genre), collocates, and concordances. In other words, users can click through the text, word by word, and get an incredible wealth of corpus-based information about any and all of these words.

Perhaps the most innovative (and hopefully useful) tool at WordAndPhrase is the ability to highlight selected phrases in the inputted text, and then have the interface suggest related phrases from COCA.

As an example, suppose that language learners had inputted the reading shown in Figure 12.6, and that they wanted to find other phrases related to *instructional methods*. The learners would simply click on these two words (*instructional* and *methods*), which are then inputted into the form below the inputted text, and they could then highlight *methods* and click [PART OF SPEECH] to find other phrases from COCA that are composed of *instructional* + NOUN: *instructional* + *strategies*, *materials*, *practices*, *time*, *methods*, *activities*, *program*, *techniques*, *programs*, *technology*.

In addition to [PART OF SPEECH], the users can select other ways to compare the phrase in their text to the 450 million words of text in COCA. For example, if the inputted text has the phrase *vintage cars*, the learners could then highlight this phrase in the text, select [SYNONYMS], and then see phrases like *old cars* (224 tokens), *classic cars* (86), or *antique cars* (52). Or—to return to an example shown above—if the students are writing a paper and they write the phrase *potent argument*, they could highlight that phrase in the paper, click on [SYNONYMS], and see the frequency of related phrases in COCA: *strong argument* (138 tokens), *powerful argument* (81), *convincing argument* (81), *persuasive argument* (63), *effective argument* (18), and so on.

The advantage of this interface over the regular COCA interface should be quite obvious. In COCA, the language learners have to input bits and pieces of the entire paper or article—phrase by phrase—and then see the related phrases for each one of these phrases—one by one. In the WordAndPhrase interface, on the other hand, they can input the entire text once. The students can then click on a phrase that they would like to explore and compare in COCA, then

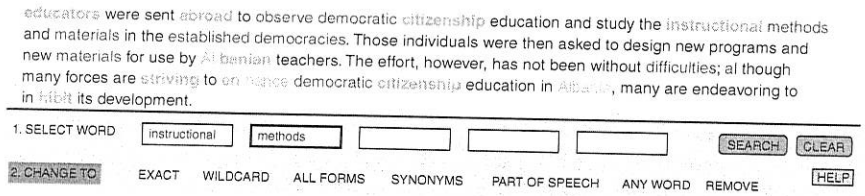

FIGURE 12.6 Selecting Phrases from the Inputted Text.

select another one, and so on. It is much quicker and easier than using COCA, and it preserves the contextual integrity of the original text. And as we have mentioned, no other tool that is currently available allows language learners to use corpus data to this extent to analyze the words and phrases of a text.

Summary

- University-level learners of English face significant challenges in terms of learning vocabulary and using vocabulary appropriately in their writing. There are a number of types of variation that make it hard to find "just the right word":
  - Context is crucial and word choice is a function of the words that are nearby (collocates).
  - Learners have to choose between (near) synonyms.
  - "The right word" is often a function of genre (e.g., formal or informal).
  - "The right word" is often a function of dialect (e.g., British vs. American).
  - Word choice is also a function of language change—some words just sound too innovative or too "old-fashioned."
- There is no dictionary or thesaurus that provides the level of detail that learners need in order to address all of the issues mentioned above, and to have their writing sound more native-like. With a corpus like COCA (and COHA and the BNC), however, learners can quickly get information on all of these different factors.
- With tools like WordAndPhrase, which allow users to input entire texts and then focus on specific words and phrases in those texts (using corpus-based data), this is made even easier.

It is worth noting that most of these tools are very recent (i.e., COCA became available in 2008; COHA in 2010; and WordAndPhrase in 2012), which suggests that language learners today have access to a number of important resources that were not available to previous learners. It is our hope that these resources will become a more integral part of literacy training for nonnative students in higher education, and that effective methods will be developed to train learners to use them in the most productive ways.

Discussion Questions

1. What are some benefits of using corpus-based data to inform the writing decisions of nonnative English learners in higher education?
2. What vocabulary support is available through corpus searches that is not available through dictionaries and thesauri?

3. After reviewing the tutorials on corpus.byu.edu/coca, perform your own collocate, synonym, and genre-based searches. What possibilities for reading and writing support do you see from your own experience using the corpus? What potential limitations do you see? What training may be needed for nonnative English writers and readers to take advantage of these powerful tools?
4. Enter a short text on WordAndPhrase (perhaps a piece of your own writing) and utilize the word and phrase functionality. What potential writing support do you see when using this corpus-based tool? What training would be required for nonnative English learners to take advantage of this tool to support their English literacy needs?
5. How can corpus-based searches inform the intuitions of language teachers?

Further Reading

Aijmer, K. (Ed.). (2009). *Corpora and language teaching*. Amsterdam: John Benjamins.

Bennett, G. R. (2010). *Using corpora in the language learning classroom: Corpus linguistic for teachers*. Ann Arbor, MI: University of Michigan Press.

Granger, S., Hung, J., & Petch-Tyson, S. (Eds.). (2002). *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins.

Reppen, R. (2010). *Using corpora in the language classroom*. Cambridge: Cambridge University Press.

References

Aijmer, K. (Ed.). (2009). *Corpora and language teaching*. Amsterdam: John Benjamins.

Bennett, G. R. (2010). *Using corpora in the language learning classroom: Corpus linguistic for teachers*. Ann Arbor, MI: University of Michigan Press.

Bernardini, S. (2004). Corpora in the classroom: An overview and some reflections on future developments. In J. M. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 15–36). Amsterdam: John Benjamins.

Biemiller, A. (1999). *Language and reading success*. Newton Upper Falls, MA: Brookline Books.

Blanpain, K., Heyvaert, L., & Laffut, A. (2008). Collex-Biz: A corpus-based lexical syllabus for business English. *ITL: International Journal of Applied Linguistics, 155*, 77–93.

Brooding. (n.d.). In *Dictionary.com's online dictionary*. Retrieved 2014 from http://dictionary.reference.com/browse/brooding?s=t.

Cobb, T. (n.d.). *The Compleat Lexical Tutor*. Retrieved 2014 from www.lextutor.ca.

Corson, D. (1997). The learning and use of academic English words. *Language Learning, 47*(4), 671–718.

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

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*, 159–190.

Davies, M. (2011). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing, 25*, 447–465.

Davies, M. (2012). Expanding horizons in historical linguistics with the 400 million word Corpus of Historical American English. *Corpora, 7*, 121–157.

Davies, M., & Gardner, D. (2010) *A frequency dictionary of American English: Word sketches, collocates, and thematic lists*. London: Routledge.

Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus-based methodologies. *English for Specific Purposes, 24*, 321–332.

Gardner, D. (2012). Technology and usage-based teaching applications. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1–7). Oxford: Wiley-Blackwell.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305–327.

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes, 6*, 319–335.

Goldenberg, C. (2008). Teaching English language learners: What the research does—and does not—say. *American Educator*, (Summer), 8–44.

Granger, S., Hung, J., & Petch-Tyson, S. (Eds.). (2002). *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins.

Hafner, C. A., & Candlin, C. N. (2007). Corpus tools as an affordance to learning in professional legal education. *English for Academic Purposes, 6*, 303–318.

Henry, A., & Roseberry, R. L. (2007). Language errors in the genre-based writing of advanced academic ESL students. *RELC, 38*(2), 171–198.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Jacobs, V. A. (2008). Adolescent literacy: Putting the crisis in context. *Harvard Educational Review, 78*(1), 7–39.

Kanoksilapatham, B. (2005). Rhetorical structure of biochemistry research articles. *English for Specific Purposes, 24*, 269–292.

Martínez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes, 28*, 183–198.

Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purposes, 25*, 235–257.

Nelson, M. (2006). Semantic associations in Business English: A corpus-based analysis. *English for Specific Purposes, 25*, 217–234.

Reppen, R. (2010). *Using corpora in the language classroom*. Cambridge: Cambridge University Press.

Simpson-Vlach, R., & Ellis, N. (2010) An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*, 487–512.

Sprawl. (n.d.). In *Dictionary.com's online dictionary*. Retrieved 2004 from http://dictionary.reference.com/browse/sprawl?s=t.

Wang, J., Liang, S., & Ge, G. (2008). Establishment of a medical academic word list. *English for Specific Purposes, 27*, 442–458.