Journal of English for Academic Purposes 12 (2013) 155–165

Contents lists available at SciVerse ScienceDirect

Journal of English for Academic Purposes

journal homepage: www.elsevier.com/locate/jeap

Google Scholar and COCA-Academic: Two very different approaches to examining academic English

Mark Davies*

4071 JFSB, Dept. Linguistics and English Language, Brigham Young University, Provo, UT 84602, USA

Keywords:
Corpus
Academic
COCA
Google Scholar
EAP

A B S T R A C T

In a recent article in the Journal of English for Academic Purposes, Brezina (2012) compares Google Scholar to the 91 million word academic component of the Corpus of Contemporary American English (COCA). In this article, I examine this comparison and show that – with the searches done correctly – COCA offers much more data than Brezina suggests. More importantly, I discuss at some length the many types of searches related to academic English which are possible with COCA but not Google Scholar, including searching for constructions (using part of speech and lemmas), comparisons between academic and non-academic genres or between different sub-genres of academic, creating frequency lists, finding collocates (to examine word meaning and usage), and carrying out semantically-oriented searches with synonyms and customized lists. Finally, I show how the new WordAndPhrase.info site provides even more user-friendly access to COCA data, including the ability to browse through large frequency lists of academic English and input and analyze entire texts. All of these COCA-based searches provide a wealth of information for teachers and learners of academic English, and while they can be done quickly and easily with COCA, all of them would be difficult or impossible in Google Scholar.

© 2013 Elsevier Ltd. All rights reserved.

In a recent paper in the *Journal of English for Academic Purposes*, Brezina (2012) compares Google Scholar to traditional corpora in terms of its usefulness for teaching and learning academic English. In particular, he compares Google Scholar to the 86 million words[1] of academic texts in the freely available, 450 million word Corpus of Contemporary American English (hereafter COCA for the entire corpus; COCA-A for the academic portion). (For an introduction to COCA, see Davies, 2009, 2011; as well as extended discussions of COCA and COCA-based data and exercises in Anderson & Corbett, 2009; Bennett, 2010; Brinton & Brinton, 2010; Folse, 2010; Lindquist, 2010; Payne, 2010; Reppen, 2010.)

In this paper, we will first briefly revisit the data that Brezina presents for COCA-A, and show that the actual data are quite different from what he provides in his study. More importantly, in the remainder of the paper we will discuss the many ways in which a full-featured corpus like COCA-A can provide useful data to teachers and learners of academic English, which are far more advanced and useful than the very simple approach taken by Google Scholar.

1. Getting the data right

In Section 3 of his paper, Brezina discusses the fact that COCA-A is 91 million words in size, which makes it about four or five times as large as any other traditional corpus (or the academic portion of any other corpus). And yet, he argues, it is still

* Tel.: +1 801 422 9168.
E-mail address: mark_davies@byu.edu.

[1] Brezina uses the 2011 version (86 million words), while in this paper we discuss the 91 million word version (2012). COCA continues to grow by 4 million words of spoken every year.

1475-1585/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jeap.2013.01.003

not big enough for some types of searches. The main evidence for this claim comes from an analysis of "reporting verbs" (e.g. *as Jones (1998) points out*, or *as Anderson and Smith (2007) explain*). He attempts to show that while such constructions are common in the billions of words of text in Google Scholar (Google Scholar), even the 91 million words in COCA-A is not sufficient to provide much data for this construction.

Unfortunately, the data from COCA-A that Brezina presents are quite incorrect, and the number of tokens that he finds is only a small fraction of what is actually available in the corpus. For example, in Table 4 of his article, Brezina claims that there are only 26 tokens of this construction with *point out* in COCA-A, but in fact there are 320 tokens (for the phrase *as [np*] [point] out*: e.g. *as Johnson points out*) – more than 10 times what he suggests.[2] He further claims that there are no tokens at all for adverb-modified constructions, such as "*as Billings* rightly *notes*", when in fact there are 28 tokens.[3] Finally, in a summary of his critique of COCA-A, Brezina claims (2012: 324) that:

> a standard reporting structure as-author-reporting verb [e.g. *as Ellis (1999) points out*] does not occur frequently enough to be subjected to detailed analysis. Although there are dozens of examples of this structure in the [COCA-A] corpus (see Table 4), there are only [a] handful of those with a specific reporting verb (*point out*).

If we are interpreting this quote correctly, Brezina suggests that very few distinct verbs in COCA-A occur with the "reporting verb" construction. But if this is in fact his claim, then this is incorrect as well. A quick search in COCA shows that there are nearly twenty different verbs[4] that occur with the construction at least fifty times (e.g. say, point out, put it, suggest, explain, note, argue, write, observe, describe, state, observe, see it, state), which should be a sufficient number of tokens for each verb, in order for teachers and students to use the examples.

It is doubtful that Brezina deliberately misrepresented the number of tokens in COCA-A. Rather, these significant undercounts of tokens in COCA-A may simply be due to inexperience in knowing how to query the corpus to retrieve the desired results. What the actual data does suggest, however, is that COCA-A – the largest traditional "corpus" of academic English – is in fact robust enough to look at the "reporting verb" construction and – as we will see – virtually every other word, phrase, and construction that might be of interest to learners and teachers of academic English. And much more significantly, we will see that COCA-A allows for an extremely wide range of searches that provide insight into academic English, virtually none of which are possible with the limited Google Scholar "corpus".

Finally, we should acknowledge Brezina's statements that he "does not mean that the Google Scholar virtual corpus would be suitable for all kinds of corpus-based analyses of EAP" and that the "Google Scholar virtual corpus should not replace other corpora of academic writing" (2012: 330). In other words, he does not claim that Google Scholar will fulfill all of the needs that we have for a corpus of academic English. In the sections that follow, we will discuss in some detail why this is the case.

2. Simple vs advanced frequency searches

As we have discussed, Brezina deals at some length with the issue of size, and how Google Scholar is much larger than COCA-A. A much more important issue than the simple number of tokens in Google Scholar and COCA-A relates to the ease in which teachers and learners can extract the data from the two "corpora". Continuing with Brezina's construction of interest – reporting verbs – teachers and learners might be interested in what this list of verbs includes in the first place – *report*, *suggest*, *explain*, *note*, *point out*, etc. In COCA-A, using the collocates function (discussed below) we can retrieve a frequency-ordered list of hundreds of different reporting verbs (nearly 6000 tokens overall) in just 3–4 s.[5] This list includes say, point out, put it, suggest, explain, note, argue, write, observe, describe, state, observe, see it, state, and so on. And then for each verb in the list, we can click to see the verb in context – with up to a paragraph of context for each token.

Using Google Scholar, it is quite impossible to retrieve a comprehensive list of reporting verbs. In COCA-A, we can search for a "verb" ([vv*]) within two words after "*as* + proper noun ([np*])", e.g. *as Smith* [np*] *notes* [vv*]. But because Google Scholar doesn't know what a "verb" or a "proper noun" is, there is no way to search for all matching verbs. We are forced to consult some other resource – perhaps a book or some other online site – to get the list of verbs, and only then can we search for each one individually in Google Scholar, which (if possible in the first place) would take several hours. In summary, Google Scholar may be able to show us snippets of text for specific words and phrases, but – unlike COCA-A – it can't search for "constructions" per se, or suggest what the most frequent words in a construction might be.

As we have seen, even those searches that should be relatively basic in Google Scholar are seriously limited because we cannot search by part of speech. In addition, Google Scholar does not allow us to search using punctuation, which is often an important element in a construction. For example, as I wrote the second sentence of this paper ("*In particular, he considers…*"), I wondered how frequent this construction actually is in English, and how its usage in academic texts compares to non-academic texts. In COCA, we can search for the phrase [_._*In_particular*_,_], where the initial period (full stop) indicates that the phrase is sentence-initial, followed by a comma.[6] There are 1858 tokens in COCA-A – easily enough tokens for use in

[2] See http://corpus.byu.edu/coca/?c=coca&q=19414367 to perform this query and see the results.
[3] See http://corpus.byu.edu/coca/?c=coca&q=19414884.
[4] See http://corpus.byu.edu/coca/?c=coca&q=19415083.
[5] See http://corpus.byu.edu/coca/?c=coca&q=19415083.
[6] See http://corpus.byu.edu/coca/?c=coca&q=19416122.

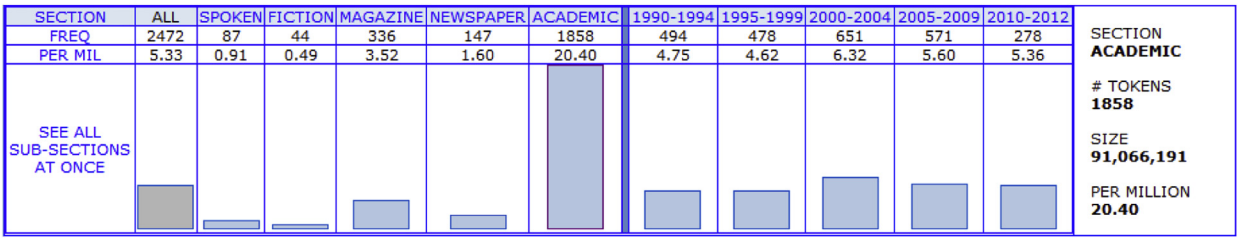


Fig. 1. Frequency in COCA by genre: *in particular.*

the classroom. In Google Scholar, on the other hand, we cannot know how many times this sentence-initial phrase occurs, since we cannot use punctuation as part of the search.

A final example shows how we can obtain frequency information from COCA-A in ways that would not be even remotely possible with Google Scholar. Suppose that a teacher wanted to discuss the use of **ly* sentence-initial adverbs in academic English. Sentential adverbs are an important and useful topic for language learners, because of the role that these adverbs play in expressing stance and the author's attitude about what is being said in the following sentence.

In COCA-A, it is quite easy to submit a query that shows the most frequent sentential adverbs in a frequency-ranked list. Users would simply input the query shown in brackets [_._*ly.[r*]_,_],[7] and they would then see a list that includes *finally* (6966 tokens), *similarly* (4184), *consequently* (3085), *additionally* (2796), *unfortunately* (2381), *specifically* (2376), *accordingly* (1351), and so on (all of which incidentally occur with more than enough frequency for use in the classroom). Google Scholar, on the other hand, cannot search for parts of speech (e.g. adverb), nor parts of words (e.g. words ending in *-ly*), nor punctuation (e.g. period before and comma after). And even if it could, it still cannot produce a frequency listing of all matching forms, as we can do in COCA-A in just a couple of seconds.

3. Comparing frequencies in academic and non-academic and in specific domains

COCA-Academic is of course just part of the larger Corpus of Contemporary American English (COCA), which contains 450 million words of text from 1990 to the present (20 million words each year). The corpus is divided evenly between the five main genres of spoken, fiction, popular magazines, newspapers, and academic texts. Unlike Google Scholar, which does not allow people to directly compare the frequency of words, phrases, and constructions in academic and non-academic English, such comparisons are both quick and easy in COCA.

For example, returning to the example of [_._*In_particular*_,_], we simply use the CHART function of COCA to see the frequency in each of the five main genres, and we see that this phrase is much more common in academic than in other genres like spoken – more than twenty times as frequent (Fig. 1).

Of course we can search for more than just simple words and phrases. For example, considering again the issue of reporting verbs, the following chart shows the frequency of [iasi[np*]i (i [m*]i) i [vv*]i] (*as* + proper noun + parenthesis + number + parenthesis + lexical verb, e.g. *as Jones (1998) notes*),[8] and we see that the construction is essentially limited to academic texts. Note that the non-academic texts are actually not even reporting verb constructions, e.g. (newspaper) *the rally fell short* as Deerfield (3-1) held *on to win* (Fig. 2).

This comparison of +academic and −academic texts can be quite useful for students, who may not be sure whether a particular word, phrase, or construction is too informal for an academic paper. For example, consider the following two charts, which show the frequency across COCA genres for *a lot of [nn*]* (e.g. *a lot of things*) and for *end up V-ing* (e.g. *he'll end up paying too much*). Both constructions are much less frequent in COCA-Academic texts, which may suggest that non-native speakers may want to avoid these in formal, academic writing. Similar charts could be given for thousands and tens of thousands of other words, phrases, and constructions (Fig. 3).

In addition to showing the frequency for particular words, phrases, or constructions in different genres, COCA can also produce lists showing what words and phrases are more common in the Academic portion of COCA, even when we do not know ahead of time what these might be. In the search form, users would simply select [vvi] "infinitival verb" for the search term (from the drop-down list) and then select [Academic] for one section of the corpus and then (for example) [Fiction] for the second section against which to compare the first section.[9] The resulting list shows that the following verbs occur much more in academic than in fiction – operationalize, remediate, predominate, aggregate, reformulate, reconceptualize, abrogate, individualize, facilitate, marginalize, and *reify* (while those that are more common in fiction than in academic include *whimper*, *snore*, *waltz*, *squeal*, *slump*, *twirl*, *perk*, *snuggle*, *quiver*, *sob*, *claw*, *saddle*, and *croak*). In terms of phrases, another

[7] See http://corpus.byu.edu/coca/?c=coca&q=19416264.

[8] See http://corpus.byu.edu/coca/?c=coca&q=19922219. Note that variations can be made on this search, such as including either parentheses or square brackets, or different forms of numbers or years.

[9] See http://corpus.byu.edu/coca/?c=coca&q=19921354.

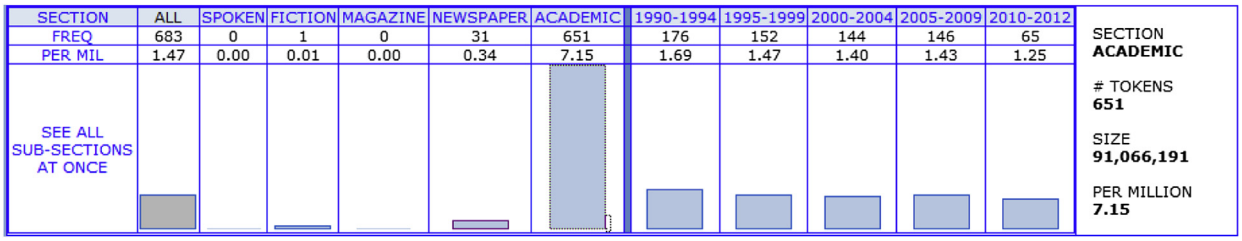


Fig. 2. Frequency in COCA by genre: reporting verbs.

simple search shows that the following phrasal verbs with *out* are much more common in academic than fiction – *bear out*, *contract out*, *phase out*, *emerge out*, *opt out*, *carry out*, *rule out*, *map out*, *separate out*, and *sketch out* (while the following are much more common in fiction: *stare out*, *freak out*, *scream out*, *pop out*, *make out*, *bust out*, *chill out*, and *chicken out*). COCA knows the frequency of each word and phrase in each genre, and we can thus compare across genres and produce lists like this in ways that would not be possible with Google Scholar.

In addition to comparing academic to other genres, COCA can also compare domains and disciplines *within* the academic genre. For example, the following charts show the frequency of the word *diagnosis* and the lemma *regulation* in the different domains of academic (Figs. 4 and 5).

And as before – when we used COCA to generate lists of words that are more common in academic than in non-academic texts, we can also use the corpus to generate lists of words that are more frequent in one particular domain of academic. For example, in just a few seconds we can generate a list of adjectives that are much more common in medical texts than in other domains of academic, such as *endoscopic* (486 tokens in Academic:Medicine), *parotid* (391 tokens), *preoperative* (579), *postoperative* (1270), *tympanic* (340), *orthopedic* (540), *mucosal* (300), *vestibular* (405), *venous* (416), *nasal* (1943), *renal* (951), *cervical* (563), *urinary* (306), and *arterial* (323).[10] Likewise, we can generate a list of verbs that are much more common in religion and philosophy, such as *preach* (85), *worship* (82), *pray* (139), *affirm* (139), *heal* (112), *suppose* (118), *confess* (55), *proclaim* (58), *love* (236), *conceive* (125), *condemn* (84), *endorse* (111), *correlate* (113), and *transcend* (91).[11]

While COCA can easily generate such lists – either between academic and non-academic texts, or for particular domains of academic – this would be very difficult or impossible in Google Scholar. As far as the first type of list – academic vs. non-academic, there is really no way to generate such lists. As far as domains go, neither can Google Scholar automatically generate a list of all words that are more frequent in one domain than another. The best it can do is show the frequency of a given word in different domains, but even here it is quite cumbersome. We would need to enter each word individually into Google Scholar, select a particular domain, write down the frequency, search by another domain, and continue until we had done individual searches in each domain. If we wanted to check another word, we would then start over again, and it would take hours for a list of 100–200 words. In COCA-A, we can generate such lists in a matter of 2 or 3 s.

4. Word and phrase patterns

Of course there is much more to a word than knowing its frequency. We also want to see *how* it is used – the patterns in which it occurs and the other words with which it occurs, which can provide valuable insight into the meaning of the word. Let us look first at patterns. In COCA, it is quite easy to generate re-sortable concordance lines for a given word or phrase. For example, the following are a handful of lines for the word *diametrically* in COCA-Academic texts. (Note that for reasons of space in this printed article, the number of words per line has been significantly reduced, and the words that are color-coded for part of speech in the web interface appear in grayscale here.) (Table 1).

Of the 100 sample lines from COCA-A, approximately 95 occur in the set phrase *diametrically opposed to the*. This type of information may appear in a good learner dictionary, but some word patterns are somewhat more complicated. Take for example the word *arguably*, as seen in the following selection of concordance lines from COCA. As we see here, *arguably* usually occurs in the context of hedging and softening of comparisons, and it is doubtful that many learner dictionaries would or could explain this as well as the concordance lines (Table 2).

Finally, the pattern of some words is even more complex. Take for example the word *budge*, as shown in Table 3.

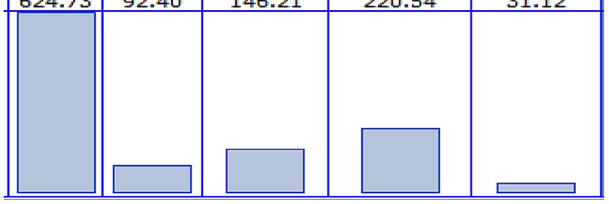
As we can see, *budge* nearly always occurs in a context in which it is preceded by a negative word, or at least negative emotions. This is useful information for a non-native student who is wondering whether it could be used in a sentence like "*he budged, and let the process proceed forward*" (it would be awkward), but it is the type of information that is often not found in a dictionary (although some advanced learners dictionaries may at times include such information).

[10] See http://corpus.byu.edu/coca/?c=coca&q=19423594. Users select [ACAD: Medicine] for the first section and [Academic] (General) for the section against which to compare it, and then select [Adjective] ([jj*]) from the drop-down list of parts of speech. Note that the words in this list are ranked not by raw frequency, but rather by how more frequent they are in medical texts than in non-medical texts.

[11] See http://corpus.byu.edu/coca/?c=coca&q=19423673 and the explanation in the preceding note regarding the ordering of words.

a lot of [nn*]

SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
59702	8356	13972	20227	2834
624.73	92.40	146.21	220.54	31.12



end up [vvg*]

SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
2796	1316	2475	2066	711
29.26	14.55	25.90	22.53	7.81

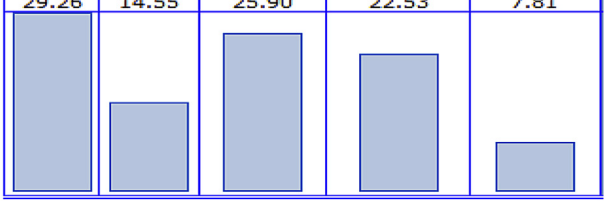


Fig. 3. Frequency in COCA by genre: two constructions.

As Brezina (2012) notes, it would be quite difficult to generate re-sortable concordance lines like this with Google Scholar. We would have to copy several pages of results from Google Scholar to another program, clean up and format the text (not a trivial task), and then import the data into a program like MonoConc, WordSmith, or AntConc, and finally display the concordance lines there. This process, once we have mastered the learning curve, would take perhaps 2 or 3 min per word. But with COCA, we can do it in 1–2 s.

5. Using collocates to examine word meaning and usage

Another important tool for examining word meaning and usage are collocates, or "nearby" words. With COCA, it is very easy and quick to extract collocates. We simply enter the word, click on the "collocates" button, and then (optionally) specify the part of speech of the collocate and the "span" of words in which the collocate appears (e.g. two words to the left of the node word, or four words to the right). For example, users can search for the nouns occurring nearby after the verb *break*, and in less than 2 s they would see a list like *law* (1527 tokens), *heart* (1454), *news* (1357), *record* (995), *rules* (943), *silence* (896), *ground* (804), *leg* (567), *barriers* (486), *cycle* (468), and *pieces* (445).[12] They 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" (actually the Mutual Information score[13]), 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 & Garder, 2010 for the most frequent collocates of the top 5000 words in American English.)

Consider a somewhat more interesting 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* US, that it is associated with *pollution* and *congestion*, and that people are trying to *reduce*, *stop*, and *fight* against it. As we can see, collocates "paint a picture" of a word that is far beyond what virtually any dictionary can provide.

The preceding examples are from the entire COCA corpus (450 million words; all genres), but we can of course limit the collocates to just those occurring in the 91 million words of academic texts. And it is important that we be able to limit the searches to search just academic, because there are of course important differences in the meaning and usage of a word between academic and non-academic texts. Fortunately, in COCA we can compare the collocates of any word or phrase in academic and non-academic texts. To take a somewhat trivial example, the most common collocates of *chair* in academic texts (compared to non-academic) are program, faculty, member, role, education, design, research, commission, dean, leadership, subcommittee, science, community (=head of a committee), while the most common in non-academic texts are edge, mother, wing, feet, kitchen, foot, beach, captain, eyes, lawn, canvas, leather, reading, hands, plastic, arms (=physical chair, to sit in).[14] For some words, the difference in collocates is somewhat more subtle. For example, the collocates of *chain* in academic are commodity, value, analysis, covenant, realist, migration, production, supply, information, product, management, distribution, system, levels (=more abstract meaning of *chain*), while in non-academic texts they are neck, door, gold, man, cross, watch, head, fingers, padlock, medallion, fence, pocket, ankle, dog, leather, silver, locket, necklace (=a literal, physical *chain*).

[12] See http://corpus.byu.edu/coca/?c=coca&q=19921524. In this case, we have chosen a "span" of four words to the right of *break*, meaning that the noun collocate could occur within one word (e.g. *break barriers*) or up to four words (e.g. *break all of the rules*). Much beyond four words, we might start getting words that don't relate to *break*, e.g. the word *marriage* in *broke the news about the marriage*.

[13] See http://wordbanks.harpercollins.co.uk/Docs/Help/statistics.html for an overview of the Mutual Information score, and http://corpus.byu.edu/mutualinformation.asp for a discussion of how it is calculated in the BYU corpus interface.

[14] See http://corpus.byu.edu/coca/?c=coca&q=19446252.

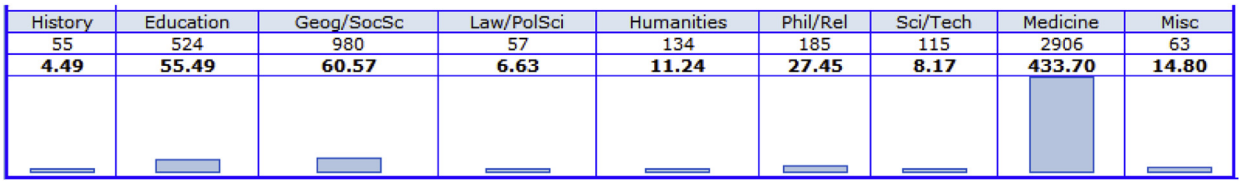


Fig. 4. Frequency in COCA-A by sub-genre (Medicine): *diagnosis*.

And of course we can limit our collocates searches just to the 91 million words of academic in COCA. For example, we can quickly and easily find the collocates of *probable*, which include (noun) *cause*, *standard*, *consequence*, *outcome*, *explanation*, *reason*, *symbol*, *scenario*, *requirement* (non-noun) *most*, *seem*, *highly*, *possible*, *criminal*, *administrative*, *less*, *definite*, *determine*, *quite*, *equally*. Likewise we can find the collocates of *significance*: (adjective) statistical, historical, special, practical, particular, religious, symbolic, spiritual, clinical, deep (noun) *level*, *test*, *meaning*, *regression*, *testing*, *coefficient*, *magnitude*, *5%*, (verb) *reach*, *approach*, *test*, *attach*, *lie*, *appreciate*, *grasp*, *imbue*, *underscore*, *accord*.

Finally, with COCA we can compare the collocates of two different words (in just academic or in the entire corpus), to see subtle differences in the meaning and usage of the two words. To take just one example, students might be confused about the difference between *significant* and *important*. The COCA-Academic texts show that the collocates of *significant* are decrease, amounts, variance, correlations, disabilities, reductions, coefficients, correlation, percentage, decreases, amount, decline, proportion, and *increase* (many of these are related to measurement and quantifying data), while those of *important* are ingredient, lessons, tools, motives, essay, key, right, ally, materials, partner, competencies, feedback, mechanisms, consideration, message, and *deadlines*. For example, *significant decrease* is much more frequent in the corpus than *important decrease*, and *important key* is much better than *significant key*. These are things that seem obvious to native speakers of English, but for non-native speakers, it is very useful to see the contrasting lists of collocates from the corpus, to help determine which word would sound more natural.

Again, it would be extremely difficult to find collocates of a given word or phrase in Google Scholar. As with the concordance lines, we would have to download several web pages of links and snippets, clean and process them, and then use a separate program to extract the collocates – all of which might take several minutes for each word (unless we hired someone to write a program to do this for us automatically). But even with such a program, we would only look at the first 1000 (random) occurrences of the word or phrase (the maximum allowable from Google Scholar), rather than the collocates with *all* tokens of the word. Such a sparse sampling would likely produce a very weak and uninformative list of collocates for the word. In COCA, on the other hand, we can easily extract the collocates of tens or hundreds of thousands of tokens of a word or phrase in just 2–3 s. And as we have seen, we can also quickly and easily compare collocates of different words to see differences in meanings, which would be quite impossible with the simple Google Search "corpus".

6. Using synonyms for more advanced semantically-oriented searches

With COCA, it is possible to do even more advanced semantically-oriented queries, using the built-in "synonyms" component of the corpus. This can be very useful for learners and teachers, to help them search for semantic concepts, rather than just strings of words (as with Google Scholar).

For example, suppose that a student wants to see the synonyms of *precarious*. She would simply enter [=precarious] in the search form, and she would then see something like the following table, which shows the frequency of each synonym and its frequency in each genre. The fact that *precarious* is only used about one tenth as much as *weak* or about one sixth as much as *slight* suggest to the learner that this word has a much more narrow 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 that follows.) (Table 4).

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 click on another synonym, and so on. For example, if a student clicks on *precarious*, she would then see *dangerous*, *uncertain risky*, *hazardous*, *unstable*, *shaky*, *unsafe* (among others), and she could then click on *shaky* to see *uncertain*, *trembling*, *questionable*, *unstable*, *dubious*, *doubtful*, *unreliable*, and so on. In this way,

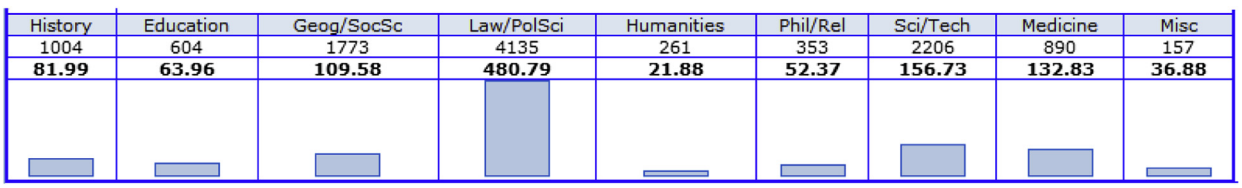


Fig. 5. Frequency in COCA-A by sub-genre (Law/PolSci): *regulation*.

Table 1
Concordance lines in COCA-A: *diametrically.*

2000	ArtBulletin	caricature") could be seen as	diametrically	**opposed** to Rivera's tendency to
1996	IntlAffairs	determination , therefore , are	diametrically	**opposed** to that of Catholic and Muslim
2010	LawPublPol	mother." These findings were	diametrically	**opposed** to the state of affairs that
1999	ChurchState	in 1997 , and most were	diametrically	**opposed** to the formal legal situation
2001	ClearHouse	directions for reform that are	diametrically	**opposed** to the "solutions" advocated
1996	EnvirAffairs	This approach is	diametrically	**opposed** to the former EPA analysis

she can follow through the chain of meaning, from one sense to another. And in each case, she could see the frequency of the synonyms and their distribution by genre, to 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.

As we saw with collocates (e.g. the collocates of *chair* or *chain*), the frequency of synonyms can vary greatly from one genre to another, and this can be quite important if we are interested primarily in academic English. Suppose, for example, a student wants to see which synonyms of *strong* are used in different genres. With one simple search, she could see which synonyms of *strong* are more much common in academic writing (e.g. effective, deep-seated, clear-cut, durable, compelling, robust, persuasive, dedicated, potent, powerful), and which are more common in informal, non-academic genres like fiction (e.g. *beefy*, *burly*, *strapping*, *spicy*, *pungent*, *brawny*, *well-built*, *biting*, *sturdy*, *dazzling*). 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 of course any genre!), or expect to see *deep-seated hands* or *compelling wind* in a short story (or again, any genre). This is obviously a real improvement over simple thesauruses, which simply give lists of synonyms, but little or no indication of which items are more frequent in different genres.

The synonym feature is the most useful for learners when the synonym is part of a particular phrase. For example, suppose that a student is writing an academic paper and that he is considering using the phrase *potent argument*, but he wants to see whether there are better, more common ways to express this. He would simply enter [=*potent*] *argument* into the search form, and he 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 with a given word (and in a given genre), which is something that even the best of thesauruses could not do.

Again, most of what we have described regarding synonym-based searches in COCA would be either impossible or very cumbersome in Google Books. As far as finding the frequency of different synonyms, one could conceivably consult a thesaurus and enter each word into Google Scholar, and then copy these frequencies over to a spreadsheet, for example. To then compare different synonyms in a phrase with a given word (as in the example of *potent argument* above), they would search for each of these 15–20 phrases – one by one – and see which is more common. But with COCA, the same student could do this in 1 or 2 s.

7. A more learner-friendly interface: www.wordandphrase.info

As nice and useful as COCA is 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 a learner has a 500–600 word paper that she has written, she

Table 2
Concordance lines in COCA-A: *arguably.*

1990	HlthSocW	implications of their decision , it	**is**	arguably	better practice to enable people to go
2004	SocSciRev	quantities of facts . Moreover , it	**is**	arguably	much easier to tell students that water
1993	Environ	western parts of the region , it	**is**	arguably	one of the most outdated and inefficient
2011	GeogRev	Analysis of Chinese Geomancy	**is**	arguably	one of the influential studies on the
2005	Hemisph	and market-oriented skills	**is**	arguably	the biggest single challenge to Brazilian
1997	AnthrQ	noted that in Israel , the military	**is**	arguably	the most important social network.
1994	EnvirHealth	Procurement Policies # This	**is**	arguably	the most important component of any
2002	EnvirHealth	that bioterrorism (BT) response	**is**	arguably	the most significant public-health issue
1997	InstrPsych	their thoughts and emotions . This	**is**	arguably	the most student-centered and

Table 3
Concordance lines in COCA-A: *budge*.

2006	CollStudies	window 's frame . It wo **n't**	budge	. Trying again with more force
2003	ABAJournal	to their privileges and would **n't**	budge	. Justice first put the waiver policy in
2007	ABAJournal	but on this point the Iraqis would **n't**	budge	. # So , after all the wrangling , it was a
2004	OrthoNurs	daughter refusing to eat would **never**	budge	. # In July , my parents started to look
2005	ArtBulletin	herd . When he finds he can **not**	budge	the thing from the ground , the local
1993	CrossCurrent	their action . The board did **not**	budge	, and Mixter and the rest of the faculty,
2003	PubInterest	began , the onset of fighting did **not**	budge	the war 's strongest opponents . This

would need to copy and paste many individual "snippets" from the paper (e.g. 1–5 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 bit of a learning curve in terms of 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 importantly, it also allows users to enter and analyze entire texts, rather than having them enter many individual words and phrases, as with the regular COCA interface. In the following section, we will discuss the ability to enter entire texts into WordAndPhrase. First, however, we will briefly examine how WordAndPhrase provides information on individual words.

Via the www.WordAndPhrase.info 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 Fig. 1). 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 domains 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 – at www.WordAndPhrase.info all of this information is provided at one time. Basically anything that COCA can tell us about the word is all displayed together, with extensive links from one word to another (Fig. 6).

The second part of the site allows users to input a text of their choice (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 Fig. 2 (a random news release from Science Daily: http://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 to 3000 in the core. A final 20% of the words are not in the top 3000 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 as in the "non-academic" genres (e.g. fiction and newspapers). (Note that in the printed grayscale version below, it is difficult to distinguish core academic from domain specific words, but the web interface allows us to use color-coding to make these distinctions much more transparent.) (Fig. 7).

Table 4
Synonyms of *precarious*, by genre.

SYNONYM	TOTAL	SPOK	FIC	MAG	NEWS	ACAD
WEAK [S]	16176	2190	3816	3266	2820	4084
SLIGHT [S]	9169	796	3576	2040	1293	1464
DELICATE [S]	8377	714	2932	2333	1476	922
FRAGILE [S]	5916	738	1549	1496	1060	1073
UNSTABLE [S]	3097	458	339	670	390	1240
SHAKY [S]	2493	323	856	523	609	182
FRAIL [S]	1986	188	898	347	244	309
BRITTLE [S]	1679	79	730	491	234	145
PRECARIOUS [S]	1612	182	299	334	280	517
TENUOUS [S]	1364	126	219	266	255	498

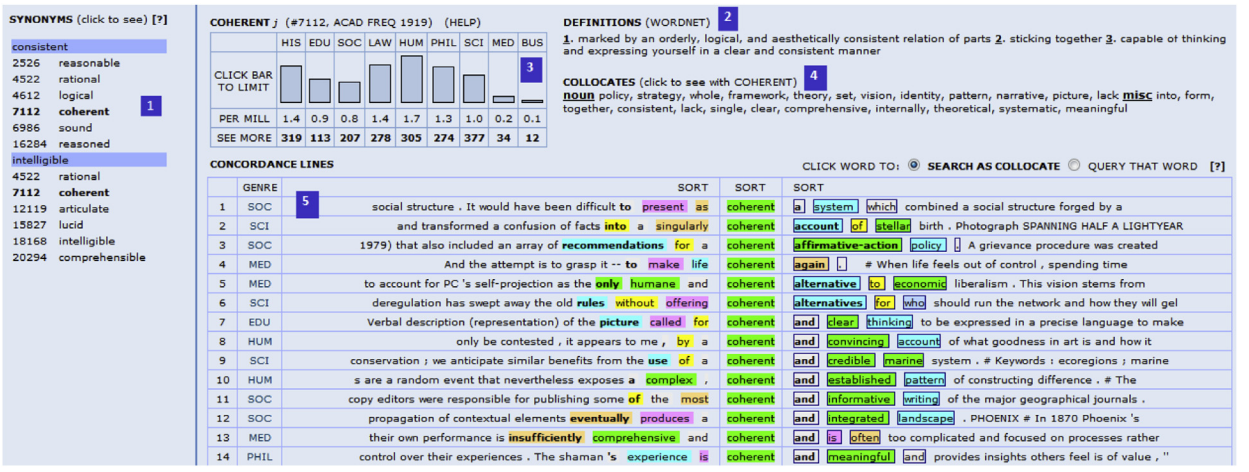


Fig. 6. www.WordAndPhrase.info: *coherent*.

At the most basic level, users can click on the different frequency bands (see the top of the figure: lemmas 1–500, lemmas 501–3000, lemmas above 3000, and also "academic"), to see the words from those lists. For example, by clicking on [>3000], a student would see the following words, which provide a fairly nice summary of what the article is about (Table 5):

Other sites, such as Tom Cobb's Compleat Lexical Tutor (http://www.lextutor.ca) offer similar functionality. But www.WordAndPhrase.info offers something quite useful, which is (to our knowledge) not available at any other site. This is the ability to click on any of the words in either the customized lists (e.g. Table 3 above) or any of the words in the original text (Fig. 7 above), and then see the full-featured entry for that word (e.g. Fig. 6 above) – 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 www.WordAndPhrase.com is the ability to highlight selected phrases in the inputted text, and then have it suggest related phrases from COCA (Fig. 8).

As an example, suppose that the language learner had input the reading shown in Fig. 8 above, and that he wanted to find other phrases related to *instructional methods*. He would simply click on these two words, which are then inputted into the form below that, and he 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 user can select other ways to compare the phrase in his text to the 450 million words of text in COCA. For example, if the inputted text has the phrase *vintage cars*, he 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 he is writing a paper and he writes the phrase *potent argument*, he could highlight that phrase in his 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 learner has 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, he can input the entire text once. He then clicks on the phrases that he would like to explore and compare in COCA, then selects another one, and so on. It is much quicker and easier, and it preserves the contextual integrity of the original text.

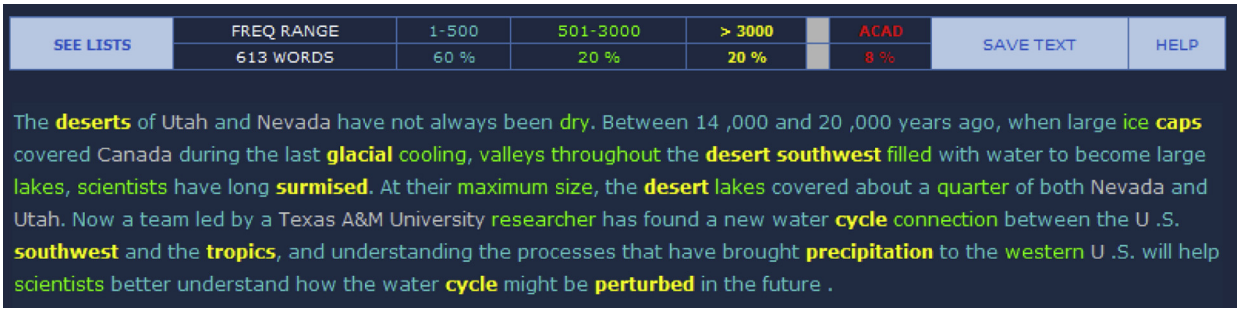


Fig. 7. www.WordAndPhrase.info: inputted text.

Table 5
www.WordAndPhrase.info: word list 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

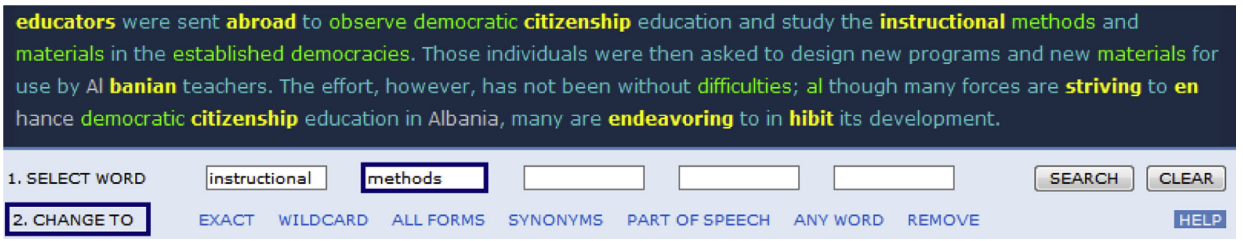


Fig. 8. www.WordAndPhrase.info: selecting phrases from the inputted text.

8. Discussion and summary

I agree strongly with Brezina that "bigger is (usually) better" in terms of corpora. For example, a comparison of COCA-Academic and the 16 million words of academic texts in the 100 million word British National Corpus shows that COCA-A, with five to six times the size of academic text – provides much richer data than the BNC. I disagree, however, with his claim that the 91 million words of text in COCA-A are too small to be of value for most words, phrases, and constructions. Undoubtedly, for some very low frequency words and phrases (such as the *fit into place* phrase that Brezina mentions), a larger corpus would be better. But as we have seen, COCA-A is clearly robust for the one primary construction (reporting verbs) where he claimed it was not.

In addition to size, there are two advantages that Google Scholar has over COCA-A and most similar "traditional" corpora. First, Google Scholar allows researchers to zero in on very specific domains, such as linguistics or even applied linguistics. While this is probably not overly useful for investigations of syntax (i.e. the grammar of linguistics is probably not appreciably different from that of other related fields), it can be very useful to see words in context in a particular domain, such as the words *morpheme* or *collocate* in linguistics articles. In addition, with Google Scholar, it is possible to limit the searches to a particular journal or (even more powerful) create a specialized corpus "on the fly" by selecting several related journals.[15]

On the other hand, as we have discussed, Google Scholar allows only the most basic types of searches in terms of words and phrases and constructions. COCA-A, on the other hand, can provide extremely useful data to teachers and learners of English on a wide range of phenomena that are typically either very cumbersome or (in most cases) impossible with Google Scholar. To summarize what we have presented in more detail in this paper, these searches include the ability to 1) compare the frequency of given words, phrases, and constructions between academic and non-academic texts, 2) generate comprehensive lists of words and phrases that are more common in academic than non-academic texts, 3) find the frequency across the different domains of academic (such as law, medicine, or education), 4) generate comprehensive lists of words that are more frequent in one of these domains than the others, 5) search by part of speech to look for constructions, 6) search using punctuation to find constructions in particular contexts, 7) provide re-sortable concordances for words and phrases, 8) search for collocates to examine meaning and usage, 9) compare collocates in academic and non-academic texts to see how the meaning in academic texts may be different, 10) compare collocates of different words (such as *significant* and *important*) to see differences in meaning and usage, 11) search for the synonyms of a given word (e.g. *strong*) to see their frequency and distribution, and 12) use the synonyms feature to find just the right word in a particular context (e.g. alternatives to "*potent*" *argument*).

In summary, while it would be either very difficult or impossible to carry out such searches with Google Scholar, all of this data on academic English is quickly and easily available to teachers and learners with the 91 million words of academic texts in the Corpus of Contemporary American English.

References

Anderson, W., & Corbett, J. (2009). *Exploring English with online corpora: An introduction*. London: Palgrave Macmillan.
Bennett, G. (2010). *Using corpora in the language learning classroom: Corpus linguistics for teachers*. Ann Arbor: University of Michigan Press.
Brezina, V. (2012). Use of Google Scholar in corpus-driven EAP research. *Journal of English for Academic Purposes, 11*, 319–331.

[15] While it would be possible to search by particular journals in COCA, we have chosen not to, for reasons of copyright and licensing, i.e. to not infringe on the domain of other resources like ProQuest Research Library or EBSCO Academic Search Premier, which are oriented toward searching specific journals.

Brinton, L., & Brinton, D. M. (2010). *The linguistic structure of modern English*. Amsterdam/Philadelphia: John Benjamins.
Davies, M. (2009). *Exploring English with online corpora: An introduction*. London: Palgrave Macmillan.
Davies, M. (2011). *Exploring English with online corpora: An introduction*. London: Palgrave Macmillan.
Davies, M., & Garder, D. (2010). *Exploring English with online corpora: An introduction*. London: Palgrave Macmillan.
Folse, K. (2010). *Clear grammar: Keys to grammar for English language learners* (2nd ed.). Ann Arbor: University of Michigan Press.
Lindquist, H. (2010). *Corpus linguistics and the description of English*. Edinburgh: Edinburgh University Press.
Payne, T. (2010). *Understanding English grammar: A linguistic introduction*. Cambridge: Cambridge University Press.
Reppen, R. (2010). *Using corpora in the language classroom*. Cambridge: Cambridge University Press.

Mark Davies is the creator of the Corpus of Contemporary American English (COCA) and several related corpora (corpus.byu.edu), which are used by 100,000+ teachers and learners every month. He has published four books and more than fifty articles on corpus design and use, and is the recipient of several large grants (three NEH and two NSF) related to corpus creation and use.