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.

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.