I’m Nigel Caplan, ESL Specialist at the University of North Carolina-Chapel Hill. This is the second video presentation on the topic of using a corpus search to improve your vocabulary. It will help to watch part 1 first if you are not familiar with corpus searching.

You will be able to follow this presentation better if you click the ZOOM icon – that’s the PLUS key above the powerpoint slide in the top right corner.
In the first video, I showed you how to use the Corpus of Contemporary American English, or COCA. The creator of that corpus, Professor Mark Davies of BYU, has also made several other corpora available online. The most important ones for our use are highlighted here: in addition to COCA, you can access the British National Corpus – a slightly smaller and older corpus that focuses mainly on British written English, and the TIME corpus, which I’ll demonstrate in a minute. All three corpora use a very similar interface, so if you can use COCA, you can use all of them. The link to the this page is under the video.
Here’s an example from the TIME corpus. This contains the full text of all articles published in TIME, a major American news magazine, from 1920 to the present. This allows you to investigate the English of journalism, which is a good guide to the way the media has chosen words over the last century. For example, I was asked recently about the use of the word “separatist” as opposed to “rebek” or “militant.” A search of the corpus, shows that the word has appeared in TIME only in certain decades, and it has had different meanings. In the 1960s, for example, the most common use of separatist was in reference to Quebec, the French-speaking province of Canada where there was a growing interest in independence at that time. More recently, the word has been used in TIME to refer to separatist movements of Serbs, Chechens, Kurds, and other ethnic groups. This kind of information can be very helpful if you’re working in political science, history, journalism, and many social sciences.
The next corpus I want to show you is a new project from the University of Michigan, called the Michigan Corpus of Upper-Level Student Papers, or MICUSP. The link is under the video. Unlike the other corpora, this one contains student writing – but only papers that got an A! There are some papers by undergraduates in their senior year, but most are by graduate students at the University of Michigan. There is a mixture of native and non-native speakers, and the papers represent a wide range of academic disciplines. You can actually read entire papers if you want.
Here’s an example of a search on MICUSP. Say I’m studying psychology and I want to know what verb tenses are possible after a phrase like “the purpose of this study.” So, I can limit the search to psychology papers and perform a simple search for the phrase. And you can see that both past and present tenses are possible. However, if you look more closely, you can see that the present tense is used in proposals and the past tense in research papers. Since all these papers received high grades, it is probably a good idea to follow their example.
Michigan also has a corpus of academic spoken English called MICASE. I won’t go into this now, but MICASE allows you to investigate the use of spoken language in academic settings including lectures, classes, discussions, and office hours. There is a lot of information on the website to help you search the corpus.
The last corpus I want to describe is your own. If you really want to know how language works in your field, you can build your own mini-corpus. You can use the free software AntConc – the link is on the screen and under the video. Just download the version for your computer, and run it. There is help available online, including tutorials and videos, but it’s quite easy to use the basic functions.
The next step is to find some text for your corpus. AntConc needs plain text files (.txt). One way to do this is to find a journal in your field from the library database – this is an example from Cell. Simply copy the text and paste it into Notepad, and then SAVE AS a .txt text file. You might need to tidy it up a little bit by removing captions and equations, but it doesn’t need to be perfect. Repeat this several times.

Another option, if you’re a graduate student, is to make a corpus of recent dissertations from your department. These are available digitally through your university library.
Now, load the text files using File ➔ Open Files menu option, and enter a search term in the box. You can do more complex searches too, but for now I’m looking at the use of “we” in a corpus of UNC dissertations from several different fields. You could just scan down the list for some good verbs to learn and use, but you can also do some more analysis.
If you have enough files, you can compare them using the Concordance Plot function. Here, you can see that some of the dissertations use “we” heavily, whereas others almost never do. So, in biochemistry (not shown here), “we” is used by these writers very frequently, but the sociology students almost never use “we.”
I’m now comparing the use of “methodology” vs “method” in the dissertation corpus. The first observation is that method is used 5 times more frequently than methodology. We can also look at the concordance lines and try to figure out how each word is used. It seems that methodology is most often used fairly generally – the methodology, our methodology, the two-stage methodology, whereas method is sometimes more specific, describing the kind of method: histogram analysis method, density functional method, Hayes method, etc.

You’d need to look at the lines more closely, and you can also click on the word in blue to see the word in its context in the full text. There’s a lot more you can do with this software, and I encourage you to play with it if you have the time.
I’m going to show you one final computer tool you can use once you have the plain text files of some articles or dissertations in your field. The Lexical Tutor website has a tool for keyword analysis. Simply copy and paste the text of an article (or several articles) into the box, or use the upload file option at the bottom, and click SUBMIT.
Here are the top results for a search using a dissertation in the field of nutrition from the School of Public Health. The words that the computer finds are those that are more frequent in the text than in a base corpus. That is, these words are likely to be key words in the text rather than general academic vocabulary. You’ll see many useful words here for talking about nutrition: supplements, dietary, promotional, dissemination, promoting, consumer, and so on. If you repeat this a few times, you should be able to make a good list of the key vocabulary for your field of study.
You can find links to all the websites I have discussed under the video and on our blog on the Vocabulary page. I hope you find some of these tools useful in your English studies.

Thank you for watching this presentation. Please take a moment to complete a quick evaluation survey. Link under the video.

I’m Nigel Caplan in the Writing Center at UNC-Chapel Hill.