

How to use a corpus: A learner's guide to searching

You really don't want to feel like you are looking for a needle in a haystack or that you end up with so much information you become overwhelmed.

Some time spent thinking before you search will be time well spent.

The guide below will help you think things through and make for a meaningful search experience!



01

Think about what you would like to learn (more) about.

Potential sources of inspiration could be:

- doubts you've experienced in your use of English (e.g. you've probably wondered whether it is appropriate to say X in English)
- words or expressions that you've read or heard that have caught your attention because you are not so familiar with them (e.g. you may have come across a specific word in that game you like or an expression in that series you've been binge watching)
- your (spoken or written) use of English has been changed or corrected (e.g. as learners, it can be a normal experience that we do not always understand the reasons why teachers have corrected us).

02

Keep a digital language diary with your language thoughts and insights for you to investigate using a corpus. You can use an app on your phone to make a note of these ideas and/or to audio record them – whatever suits you best.

A digital language diary can really help you capture your language insights and problems at the moment when they occur to you (or shortly thereafter). Reading and easily accessing this diary can help as you begin to formulate your questions (see #4) that a corpus search can help you answer.

03

Choose a corpus that suits your needs, that is, one which contains data which are largely aligned with your study purposes. For example:

- If you are an English for General Purposes (EGP) student, you could consult the British National Corpus (<https://www.english-corpora.org/bnc/>) for British English, the Contemporary Corpus of American English (<https://www.english-corpora.org/coca/>) for American English or the Strathy Corpus of Canadian English (<https://www.english-corpora.org/can/>) for the English variety represented in Canada.
- If you are an English for Academic Purposes (EAP) student, you may wish to consult the Michigan Corpus of Academic Spoken English (MICASE) (<https://quod.lib.umich.edu/cgi/c/corpus/corpus?page=home;c=micase;cc=micase>) or the Michigan Corpus of Upper-Level Student Papers (MICUSP) (<https://micusp.elicorpora.info/>).
- If you are studying English for Specific Purposes (ESP), the News on the Web Corpus (<https://www.english-corpora.org/nw/>) might be helpful for journalists and the Corpus of Global Web-based English (<https://www.english-corpora.org/glowbe/>) for digital materials developers.
- There are several other corpora freely available online, and you're likely to find one that meets your needs (or can be used as a fair proxy to them!). You can also check out the list that comes with these guides.



How to use a corpus:

A learner's guide to searching



04

Decide on a starting point for your corpus exploration based on your thoughts in #1 and notes in #2. Most corpus investigation, especially if you're using a corpus website, will require a specific word or expression for you to search in the corpus. Some potential examples are listed below.

- Learners of English as a foreign language (EFL) seem to fear prepositions! This is a topic that you can easily investigate in several corpora. For example, which prepositions do you use to say that two people are married and to say that a married person has two children? Corpora can help you see that one says s/he is married 'to' someone but married 'with' two kids. This can be different from your first language if you speak, for instance, Portuguese or Spanish as the preposition 'with' is used to refer to whom someone is married.
- Learners of English sometimes struggle to describe things and people. There seems to be a tendency to overuse the adjectives like 'good' and 'nice'. Without relying on these general adjectives, how would you positively describe a computer? Some of the adjectives would include 'advanced', 'sophisticated' and 'powerful' according to corpus findings.
- Let's suppose you've just recently learned the phrasal verb 'break up', but you're not too sure of its meaning and use. By looking at corpus examples, you'll find out that this is used to refer to stopping fights and to end a marriage or relationship.

05

Search for a word, phrase or expression.

Most basic searches will involve you typing in a word or phrase. For example, go to: <https://www.english-corpora.org/bnc/> (or search "British National Corpus- English Corpora"), find the search tab and empty text box, and type the word or phrase that is your focus.

There may be other actions that you can perform and select to refine your search, so please ensure you follow the directions by the developer of the corpus website/software that you've chosen.

If this is the first time you use the website or the software, you might find it helpful to read its guidance and explore it freely to familiarize yourself with the technology.

Once you have given the instructions to search, the website or software will search the corpus for examples of that word/phrase. When this is complete, the results of your search will be displayed. Depending on which website or corpus that you have used, this will normally be presented as a frequency list or in concordance lines which look like this:

Here are a couple more sets of concordances from COCA. The entries for *end up* show that it is often followed by a gerund (*ends up getting*), an adjective or past participle (*real warped, smeared*), or *in* or *with*.

1990	FIC	Bk:Postmortem	of these murders , like he chooses from the menu ,	ends up	getting what he can afford . Sort of strange each time
2001	NEWS	SanFrancisco	up getting voluntary manslaughter for our guy , so he	ended up	going to jail , but his sentence was lessened . I
1996	SPOK	CNN_Crossfire	. BARONE : Rollins does n't say that that money did	end up	going to Reagan , as I understand . FERRARO : We
1995	FIC	Bk:WalkingMidnight	, but she figured he 'd dodge . She 'd probably	end up	in the drink again that way , too . Either way
2006	NEWS	USAToday	about a pound of high explosives inside a suitcase that	ended up	in the luggage bay under the forward cabin where it
2006	SPOK	NPR_FreshAir	critic . I think a lot of people feel that they	ended up	in the wrong life . How did she get there ?
1990	FIC	Bk:HoleLanguage	she 's like some walking zombie . He says she could	end up	real warped from all this . " I never heard what
2005	SPOK	NBC_Dateline	at the end of the day , Nancy and Jack Satter	ended up	right where they would have been , if they 'd never
2010	SPOK	NBC_Today	I need to loosen up a little , and these kids	end up	running a little bit too wild . Dr-KAISER : I think
2008	FIC	SouthwestRev	-- though none of it was her fault -- she still	ended up	smeared with a reputation as a deadbeat . // That winter ,
1994	NEWS	CSMonitor	to migrate to the suburbs : Urban school districts	ended up	with an increasing percentage of minority students
2004	FIC	Bk:FatalError	paying with some chickenshit family scolding that 'll	end up	with my mom selling off the ranch to bail him out
2005	FIC	Bk:LoveWalkedIn	scene , but then to feel completely satisfied when they	end up	with other people . Before you get the wrong impression

<https://www.english-corpora.org/concordances.asp>



How to use a corpus: A learner's guide to searching



06

Observe how frequent the item that you have searched for is. Most of the corpora mentioned in # 3 are large ones. If you only find a handful of examples in these corpora, it may be an indicator that the word or expression that you're searching for is not very common.

Why does frequency matter?

- If you search for 'cause an impression' in the Contemporary Corpus of American English (COCA), no instances are retrieved. Given that this corpus contains more than one billion words of text, there's something suspicious in that result, especially as there are (at the moment of writing) 2,320 instances of 'an impression' in the corpus. This suggests that the verb-noun combination may be unnatural or inappropriate. The expression 'make an impression' yields a total of 360 instances, for example.
- Should you search for 'utterly useful', you'll find (only!) one instance in the COCA, which is again surprising given that there are one billion words in this corpus. This is even more surprising if we consider that the COCA totals 11,377 instances of 'utterly' and 49,495 instances of 'useful'. However, if you search for 'utterly useless', there are 121 instances in the corpus, showing words that go together.

07

Read some of the concordance lines (i.e. short snippets of text around a specific search word/expression) in order to see how it is actually used. Also, if available, you can have a look at more of the text around the word and see information about the context. Not only will this help you further understand the use of your focus word/expression, but it will provide you with access to considerable language input, which might be difficult to find elsewhere. It can also give you further ideas for corpus exploration, which would take you back to #1. (No-one ever uses a corpus only once!)

08

Notice where the examples come from, that is, observe in which types of text your search word/expression appears. For example, 'really good' and 'real good' are mostly used in spoken texts while 'respectively' is mostly found in written texts (generally of an academic nature).

09

Repeat # 5-8 using a different corpus to check whether the results are consistent or different. Irrespective of the outcome, consider why that's the case.

10

Summarize what you've learned in a learning journal. This will provide you with a chance to review everything you've learned in your corpus searches while also providing you with a record of your language discoveries.

