

## Exploring online corpora (BNC & COCA)

### A. Introduction

The word *corpus* comes from Latin, meaning 'a body', but in Linguistics it refers to a collection of spoken or written texts, stored in a database, which can be queried using corpus search software (in a similar way to Google searches).

Corpora can give you information about language which you can't easily find from other sources, including:

- (i) *Word frequency*: This indicates the number of 'hits' (occurrences) of a search word or phrase there are in the corpus. Obviously, the larger the number of hits, the more common (or natural) the language is in a particular context.
- (ii) *Genre*: A corpus is usually divided into different 'sub-corpora' (e.g. spoken, blog, TV/movies, fiction, magazine, newspaper, non-academic, academic) so you can see the frequency of your search word/phrase in different genres. For example, *gonna* (going to) appears 353,960 times in the COCA corpus (so it is clearly a very common expression), but most of the hits are from the TV/movies sub-corpus so we can assume that this is a spoken/informal expression.
- (iii) *Historical changes*: Some corpora give information on changes in language use over time. For example, in COCA, the adverb *seldom* (rarely; not often) occurs 1,704 times in texts from 1990-1994, but only 542 times in texts from 2015-2019. This suggests that its use is declining over time and it may seem rather old-fashioned to native speakers of English.
- (iv) *Collocations*: A corpus can show you words that tend to go together (or co-occur). For example, a search of COCA for collocations with '*high*' gives some interesting combinations such as *high risk*, *high time*, *high horse*, and *high life*. Learners often make mistakes in their writing by combining words that don't normally go together so using a corpus to find common collocations for a search word will make your English seem more natural.



- (vi) *Clusters*: Natural language often takes the form of recurrent clusters (groups) of words, such as ‘*It is important to note that...*’, ‘*It can be seen that...*’, or ‘*On the other hand...*’. Learning these common clusters can also help you improve your English proficiency and using a corpus is a quick way to find them.

## The English-corpora website

The English-corpora.org website (<https://www.english-corpora.org/>) provides the best, and most widely used, corpus-querying resources available (free of charge) at the present time. The familiarization tasks below illustrate some of the key features of the interface, using the British National Corpus (100 million words) or the Corpus of Contemporary American English (1 billion words).

## The interface

The English-corpora user interface has three tabs at the top of the screen: SEARCH; FREQUENCY; CONTEXT

- (i) The ‘search’ tab provides a screen where you can enter your search word/phrase and set the search parameters;

The screenshot shows the search interface for the word "dog" on the English-corpora.org website. The top navigation bar includes tabs for SEARCH, FREQUENCY, CONTEXT, and HELP. The search results are displayed in two columns.

**Search Parameters:**  
 Word: dog [POS]  
 Buttons: Find matching strings, Reset

**Sections:** Texts/Virtual Sort/Limit Options

1	IGNORE	2	IGNORE
	.....		.....
	SPOKEN		SPOKEN
	FICTION		FICTION
	MAGAZINE		MAGAZINE
	NEWSPAPER		NEWSPAPER
	NON-ACAD		NON-ACAD

**SECTIONS**

**SHOW** Determines whether the frequency is shown for each "section" of the corpus (in the case of the BNC, the genre). For example, the synonyms of *beautiful* in each section and overall.

Select a section

un-* verbs in FICTION	Past tense verb + over in SPOKEN
*ment in ACADEMIC	Synonyms of <i>smart</i> in FICTION
ADJ + track in NEWSPAPERS	Noun near <i>chair</i> in FIC
ADJ in tabloids	Nouns in advertising

(Optional) Select a second (set of) section(s) against which to compare the sections chosen above

un-* verbs in FIC vs ACAD	Past tense verb + over in SPOK vs NEWS
*ment in ACAD vs FIC	Synonyms of <i>smart</i> in FIC vs ACAD
ADJ + track in NEWS vs SPOK	Nouns near <i>chair</i> in ACAD vs FIC
ADJ in tabloids vs NEWS	Nouns in advertising vs MISC

- (ii) The 'frequency' tab displays a summary of the search results;

The screenshot shows the 'British National Corpus (BYU-BNC)' interface. The 'FREQUENCY' tab is selected. The search term is 'DOG'. The results are summarized in a table with columns for different sections: ALL, SPOKEN, FICTION, MAGAZINE, NEWSPAPER, NON-ACAD, ACADEMIC, and MISC. The 'MAGAZINE' section has the highest frequency at 150.51.

SECTION	FREQUENCY
ALL	7764
SPOKEN	132.98
FICTION	124.83
MAGAZINE	150.51
NEWSPAPER	84.46
NON-ACAD	20.31
ACADEMIC	26.61
MISC	83.18

- (iii) The 'context' tab displays the KWIC (Key Word in Context) concordance lines for the search word/phrase.

The screenshot shows the 'British National Corpus (BYU-BNC)' interface with the 'CONTEXT' tab selected. It displays 10 concordance lines for the word 'dog'. Each line includes a reference code (e.g., A74), a section code (W\_fict\_prose), and the surrounding text with the search word highlighted in green.

REF	SECTION	CONTEXT
1	A74 W_fict_prose	out for walks and teach them tricks and stuff. Me and Annie had a <b>dog</b> once I think. I ain't sure -- I think it was us and
2	A74 W_fict_prose	It was ages ago, so I've forgot. I think we had a <b>dog</b> , though. It had yellow hair and it used to swim in the sea
3	A74 W_fict_prose	I turn round to see what's up. She's calling to a little <b>dog</b> which is running after summat with his tail wagging like mad. It's great
4	A74 W_fict_prose	after summat with his tail wagging like mad. It's great to have a <b>dog</b> I reckon -- they're more fun than cats. You can take dogs out
5	ABX W_fict_prose	of here I'll tell him you've been up in the woods with a <b>dog</b> . He'll tell your Dad.' Philip walked back up the ride not
6	ABX W_fict_prose	the boy and disappeared again into the trees. He heard him whistling for his <b>dog</b> . Philip hoped he'd find his dog and the pair of them would clear
7	ABX W_fict_prose	. He heard him whistling for his dog. Philip hoped he'd find his <b>dog</b> and the pair of them would clear off. With luck, with his whistling
8	ABX W_fict_prose	were tearing across the field all bunched up together. After them was a black <b>dog</b> . A ewe and two lambs were trailing and the dog had got them marked
9	ABX W_fict_prose	them was a black dog. A ewe and two lambs were trailing and the <b>dog</b> had got them marked. It was the worst thing -- unless the old ewe
10	ABX W_fict_prose	to him.' It wasn't Caspar in the field. It was another <b>dog</b> , a black dog but it wasn't Caspar.' Philip looked at him

## B. Familiarization tasks

1. Go to the BNC page at: <https://www.english-corpora.org/bnc/>
2. Type in the search word 'dog' in the box at the top of the screen.
3. Click the 'Sections' box to get a breakdown of the results by genre (spoken, fiction, magazine, newspaper, non-academic, academic & miscellaneous).
4. Select 'Options' and choose PER MIL in the drop-down menu for DISPLAY. This will normalize your results to give hits per million words (a conventional measure in corpus linguistics), rather than the total number of hits. This is important because the sub-corpora are different sizes and therefore can't be compared directly.

5. Click on the 'Find matching strings' button to run your query.
  - ⇒ The results for the search are displayed in the 'Frequency' window, and show that the total number of hits for *dog* in the BNC is 7,764 (meaning that this word appears 7,764 times in total in this 100 million-word corpus). The results are also broken down into sections, showing the number of hits per million words for each genre: you can see, for example, that *dog* appears almost five times more frequently in the spoken sub-corpus (132.98 hits) than the academic sub-corpus (26.61 hits). The dark/light blue shading of the boxes provides a quick indication of frequency level so that you can quickly search for patterns in the data.
  
6. Click on the word DOG in the search results section to generate a list of concordance lines in the 'Context' window. As you can see in the top-left corner of the screen, this is the first page of results from a total of 78, with all of the hits listed either from spoken meetings (S\_meeting) or newspaper tabloids (W\_written\_newsp\_tabloid). Click on the number 100, next to 'Find sample':
  - ⇒ This produces a random set of concordance lines from the complete corpus; the source information to the left of the concordance lines indicates that the examples now come from a wide range of text types (e.g. W\_biography = written biography or S\_conv = spoken conversation).
  
7. Go back to the 'Search' window and select Chart from the settings at the top left of the screen, then click on 'See frequency by section'.
  - ⇒ This produces bar charts indicating the overall frequency of the word *dog* in each sub-section of the BNC. In this way, you can quickly compare between different genres (it is much more common in magazines than academic texts, for example).

8. Change the search word to *dog\**. The asterisk is called a 'wildcard' and signifies 'and anything else'. Run the search again (if there are any problems, press the 'Reset button').
  - ⇒ The results list all of the words in the corpus beginning 'dog' (dogma, dog-eared, etc.). For example, *dogmatic* is most frequent in the academic section (5.15 occurrences per million words), while *doggy* is most frequent in the spoken section (5.62 occurrences per million words).
  
9. Search for the expression *dog's dinner* in the BNC (note that the corpus has been 'tokenized' so that all the punctuation has been separated from the words around it – this means that you will need to include a space between *dog* and 's in your search). Click on DOG'S DINNER to see a KWIC (concordance) list in the CONTEXT screen.
  - ⇒ The results indicate that there are only 8 examples of *dog's dinner* in the whole of the BNC, so we know that it is not a widely used expression. The concordance lines show that there are 3 instances of the target phrase used in the literal sense of 'dog food', and 5 instances of it used idiomatically, to mean 'done badly'.

The screenshot shows the BNC search interface with the search term 'dog's dinner' entered. The results are displayed in a table with columns for ID, Source, Part of Speech, and Context. The search results are as follows:

ID	Source	Part of Speech	Context
1 F9X	W_fict_prose	A B C	'he said.' I take it that you see that architectural dog 's dinner down there as a skillfully-planned structure -- some sort of enormous palace."
2 G1D	W_fict_prose	A B C	Camille remembered the smell of dog. It mingled with the smell of dog 's dinner which was simmering casually in a huge open vat: nameless portions of meat floated
3 HTS	W_fict_prose	A B C	in Belfast when I was waiting at the bus stop like a fresh dog 's dinner to be carried off to Dothegirls Academy in me big grey interlocks with double gusset
4 CH5	W_newsp_tabloid	A B C	# Lap of luxury # Recession bites hard -- but NOT into the dog 's dinner # WE WANT TO GIVE PETS FOOD WE LOVE' # WHAT do you
5 AHK	W_newsp_brdsht_nat_misc	A B C	Leicester DAVID PEARS and John Liley mopped up all the points in a dog 's dinner of a Pilkington Cup semi-final at the Stoop. As Pears, the faithful England
6 B7G	W_non_ac_nat_science	A B C	. Europe's scientists are claiming that observation of the volcano is a dog 's dinner. They say the observers are underfunded and disorganised. The result is both a
7 A6A	W_misc	A B C	on pop's more adventurous independent fringes, while RM remains a likeable dog 's dinner, differing from the others in its A4 format, its glossy colour pages,
8 HRT	W_misc	A B C	sprayed down through the tower. # INVESTING 1 MILLION IN A TV DOG 'S DINNER # (----) is currently investing 1 million in a novel television advertising and sampling

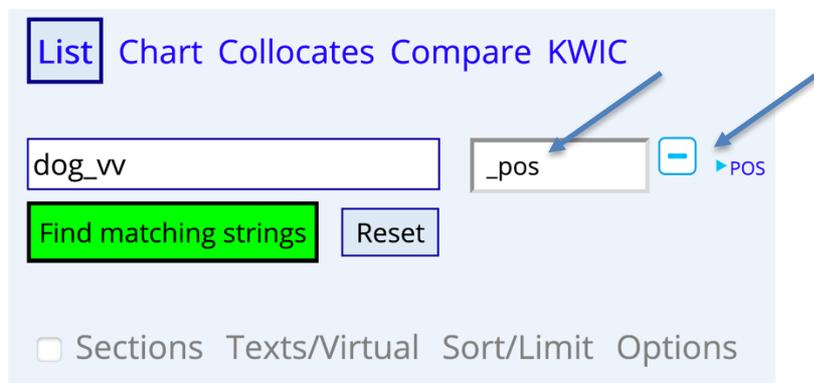
10. Change the search expression to *dog \* dog*. This time, the asterisk stands for 'any other word' since there are spaces around it.
  - ⇒ The results show *dog eat dog* is the most common pattern, unsurprisingly.

11. Return to the 'Search' window and clear the search box. Click on POS (part of speech) next to the search box and select adj.ALL from the drop-down menu (meaning all adjectives), then retype in *dog* in the box (after ADJ). Click on 'Find matching strings'.

⇒ The results show the most common adjectives used to describe dogs, with noticeable variations in frequency for different adjectives and genres. For example, 'mad dog' is much more common in newspapers than any other genre:

British National Corpus (BYU-BNC)										
SEARCH		FREQUENCY			CONTEXT				HELP	
SEE CONTEXT: CLICK ON WORD (ALL SECTIONS), NUMBER (ONE SECTION), OR [CONTEXT] (SELECT) [HELP...]										
	CONTEXT	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	NON-ACAD	ACADEMIC	COMPARE	
1	<input type="checkbox"/> LITTLE DOG	88	4.12	2.01	0.69	0.19	0.24			0.19
2	<input type="checkbox"/> BLACK DOG	77	0.20	2.39	1.79	0.96	0.30	0.20		0.29
3	<input type="checkbox"/> TOP DOG	54	0.50	0.31	3.03	0.76	0.18	0.07		0.48
4	<input type="checkbox"/> OLD DOG	52	0.60	1.57	0.41	0.29	0.06	0.07		0.62
5	<input type="checkbox"/> BIG DOG	46	1.41	1.70	0.28			0.07		0.10
6	<input type="checkbox"/> GOOD DOG	43	2.81	0.31	0.28	0.29				0.24
7	<input type="checkbox"/> MAD DOG	39	0.20	0.50	0.14	0.86	0.06			0.86
8	<input type="checkbox"/> HOT DOG	32	0.30	0.50	0.41	0.57	0.18	0.07		0.38

12. Return to the 'Search' window. Retype in the word *dog*. Click on POS (part of speech) and select `_pos`, then select verb.ALL from the drop-down menu (meaning all verbs). The search word will now be for verb forms (`_vv`) of *dog* only:



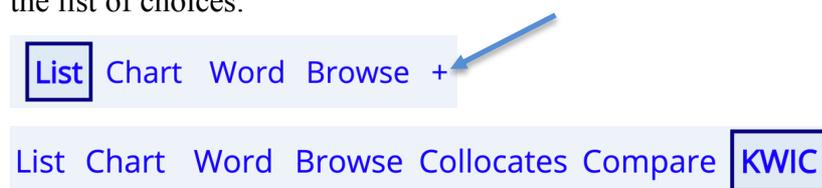
Click on 'Find matching strings'.

⇒ The results show all instances of *dog* used as a verb, with only 22 hits occurring in the whole of the BNC (so we know it isn't very common). Click on the word 'dog' to generate concordance lines in the 'Context' window. The concordance lines illustrate the two uses of the verb *dog*: (a) to follow somebody closely (e.g. How could I even think of it when you dog my every step); and (b) to cause trouble for a long time (e.g. once again injuries are beginning to dog us). Notice that the automatic POS-tagger used in the BNC has misclassified some of the hits as verbs (e.g. when you look after a police dog it becomes your pet as well). This is a useful reminder that the results are never 100% accurate!

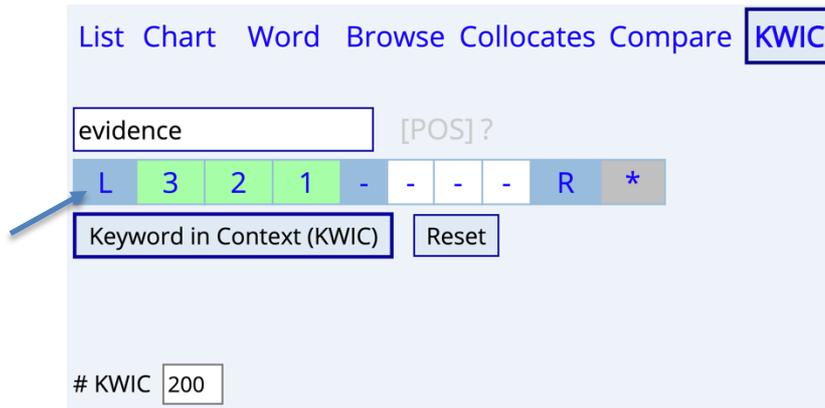
13. Click on the  icon at the top of the screen and select 'Re-do last search' (top-left column) and COCA from the choice of selections.

⇒ The results show the same search run on COCA. As you can see, there are now 430 hits for the verb *dog*, rather than just 22 – this highlights the advantages of using a larger corpus when analyzing low frequency words.

14. Return to the 'Search' window and select KWIC (key word in context) from the top left of the screen. You will need to click on the '+' symbol to expand the list of choices:



Type in *evidence* as the search word. Click on the 'L' in the 'Sort' section to show that you would like to sort the words to the left of the node (the boxes turn green to show you which side of the node is being sorted):



Then click on 'Key word in Context (KWIC)'.

⇒ The concordance lines appear in the 'Context' window, sorted to the left of the search word *evidence* (the node). The words immediately to the left and right of the node are colour coded to show word type (purple = verbs; green = adjectives, etc.). Left sorting highlights some common adjectives (*convincing, empirical, further, insufficient*) and common verbs (*give, show, find*) used with *evidence*.

15. At the top-right of the screen, select 'R' and then 'Re-sort' to arrange the concordance lines alphabetically to the *right* of the node.

⇒ The concordance lines now appear sorted to the right of the search word *evidence* (the node). *Evidence for/of/that...* now appear as common patterns in the data.

16. Return to the 'Search' window and select 'Collocates' (*collocates* are words which like to go together) from the top left of the screen. Click on 'Find collocates'

- The search results, appearing in the 'WORD' window, show the most common nouns (e.g. *piece*), adjectives (e.g. *scientific*), verbs (e.g. *provide*) and adverbs (e.g. *overwhelmingly*) which collocate with *evidence*.

+ NOUN	NEW WORD	?	+ ADJ	NEW WORD	?	+ VERB	NEW WORD	?	+ ADV	NEW WORD	?
2037	2.87	piece	2466	4.83	scientific	5455	3.37	provide	83	4.32	eg
1261	4.36	dna	2138	3.66	physical	4738	3.92	suggest	73	2.96	ie
1033	3.11	lack	2052	2.66	strong	4195	3.78	support	56	2.55	overwhelmingly
964	3.04	claim	1842	6.73	empirical	2789	4.02	present	49	4.79	conclusively
803	5.64	contrary	1665	2.56	clear	2239	2.91	base	36	2.63	scientifically
563	3.67	absence	1512	2.71	available	1478	3.22	indicate	30	3.00	improperly
510	2.61	witness	1268	3.17	far	1089	3.58	gather	21	2.78	definitively
483	3.00	existence	1243	8.04	anecdotal	987	2.58	exist	17	2.68	willfully
474	3.11	prosecutor	1062	8.24	circumstantial	887	3.08	collect	13	2.52	precious
466	2.60	jury	962	5.10	overwhelming	767	3.23	link	6	4.00	symmetrically

17. Return to the 'Search' window and select 'Compare' at the top left of the screen - two search boxes will appear below it: Word1 and Word2. Type in the search words *big* and *large* in order to compare the common collocations for these two items, then click 'Compare words'. Notice the 'Collocates' box now has an asterisk in it and the numbers 1234 are selected to indicate that the search is for any common collocates which occur up to 4 places left or right of the node.

⇒ The search results, appearing in the 'Frequency' window, suggest that *big* is used in more informal registers (e.g. *big hug*, *big mama*), while *large* is used in more formal (e.g. *large quantities*, *large samples*) and also in cooking contexts (e.g. *large saucepan*, *large eggs*).

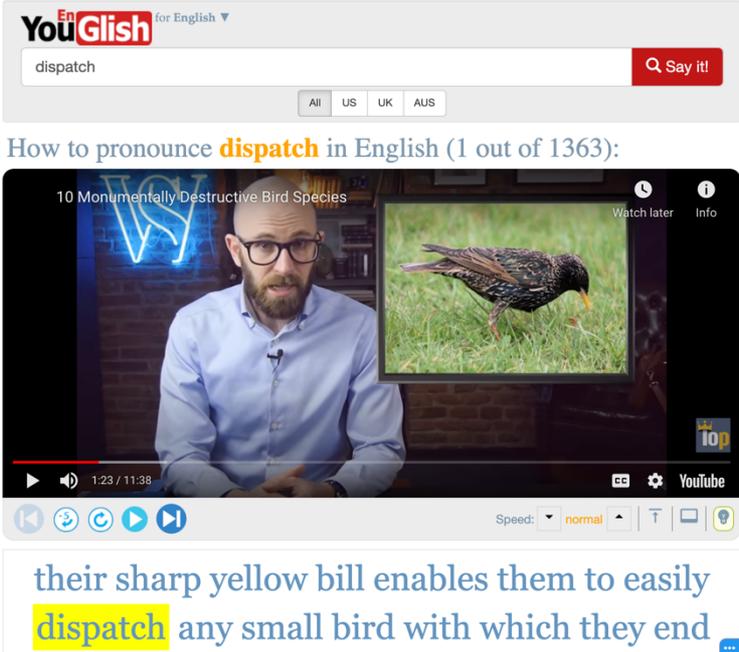
18. Return to the 'Search' window in COCA and select 'Word' at the top left of the screen. This is a very useful function which allows you to do detailed investigations into your search word. Type in the search word *dispatch* and click 'See detailed info for word'. You can see that the verb *dispatch*:

- mainly occurs in magazine and newspaper genres
- has 3 main meanings (send off promptly; complete or carry out; kill intentionally)
- Synonyms include *kill* and *send off*
- Common 2-word clusters include '*dispatched to*' and '*quickly dispatched*'

The screenshot shows the COCA interface with the search results for the word "dispatch". The interface includes a navigation bar with tabs for SEARCH, WORD, CONTEXT, and OVERVIEW. The "WORD" tab is active, displaying the following information:

- dispatch** (VERB) See: NOUN #8000
- Frequency:** A bar chart showing the frequency of "dispatch" across different genres: BLOG, WEB, TV/M, SPOK, FIC, MAG, NEWS, and ACAD. The highest frequencies are in MAG and NEWS.
- Definitions:** 1. send off promptly 2. complete or carry out 3. kill intentionally and with premeditation
- Translations:** YouGlish, PlayPhrase, Yarn, JA: Google, WordRef, Reverso, Linguee
- SYNONYMS (more):** kill, destroy, dispatch, kill, murder, slaughter, slay, send off, forward, mail, post, remit, ship, transmit
- CLUSTERS (more):** dispatch • dispatched to • dispatched by • dispatched from • dispatched with • dispatched in • dispatched him • dispatched on • dispatch photo • dispatch • has dispatched • he dispatched • had dispatched • by dispatching • have dispatched • will dispatch • quickly dispatched • they dispatched • dispatch • • dispatch photo by • dispatched a team • dispatched him to • dispatched to help • dispatched to take • dispatched to find • dispatching a team • dispatch an
- TOPICS (more):** dispatch, surrender, dispatcher, subsequently, ambulance, depart, imperial, confrontation, debris, naval, raid, scream, bomber, casualty, coordinate, fleet, pipeline, plunge, ruler, urgent
- COLLOCATES (more):**
  - NOUN: team, police, troop, officer, unit, force, us, agent
  - VERB: investigate, rescue, assist, interview, scout, copy, execute, guard
  - ADJ: remote, naval, persian, marine, armed, investigative, would-be, csi
  - ADV: quickly, immediately, easily, promptly, swiftly, secretly, abroad, quietly

- You can also link to *YouGlish*, *Playphrase* and *Yarn* to see how your search word is used in film or video clips:



The screenshot shows the YouGlish website interface. At the top, there is a search bar with the word "dispatch" entered. Below the search bar are buttons for "All", "US", "UK", and "AUS". The main content area displays a video player with the title "10 Monumentally Destructive Bird Species". The video shows a man in a blue shirt speaking, and an inset image of a starling bird. A subtitle at the bottom of the video reads: "their sharp yellow bill enables them to easily dispatch any small bird with which they end".

- You can also see a translation into your 1<sup>st</sup> language in Google translate, etc.

### C. Practice activities

1. Below are some genuine mistakes from students' essays – use the English-corpora website to identify the problem and find a more natural expression.

- (i) 'Since then, he started to go...'
- (ii) '...but we cannot make it worth.'
- (iii) 'My confidence changed...'
- (iv) '... and she died for a car accident'

\*\*\*For suggested answers, see Appendix 1 in: Gilmore, A. (2009). Using on-line corpora to develop students' writing skills. *English Language Teaching Journal* 63/4: 363-372\*\*\*

2. In the Thesaurus worksheet, we saw a video clip of a student searching for synonyms of 'fundamental': *abecedarian*, *basal*, *basic*, *beginning*, *elemental*, *essential*, *introductory*, *meat-and-potatoes*, *rudimental*, *rudimentary*, *underlying*. Use the BNC or COCA corpus to further investigate these

possible choices and decide which option is most appropriate for an academic essay.

3. Analyze some of the language you have used in one of your own essays and decide whether is natural and appropriate for an academic text.

Now you are more familiar with the corpus architecture for the English-corpora.org website, you're ready to begin exploring independently ~ good luck!

---

, but I know the manager." I can't believe my **good luck** sometimes,' he said to her later in bed.' You're always  
this season -- further details will be available from her in the Autumn. **Good Luck** with your enrolment and the start of the new academic year -- see you in  
strength of character to murder me. A chip off the old block. **Good luck**, by the way. I seem to recall I said that. Will say  
# All over India the right-angled Swastika is commonly regarded as a sign of **good luck**. Good luck is related to the literal translation of 'Swastika' which is  
On the first night Rose Lipman came backstage as usual to wish the cast **good luck**. Bunny complained of a fearful draught coming from the front of the house.  
(SP:PS1GF) (unclear) oh see you later mate (SP:KDAPSUNK) see you later, yeah, **good luck** to you (SP:PS1GF) where, where you off to? (SP:PS1GE) ta la mate (SP:PS1GF)  
. One minute to go and the Director wishes everyone down on the floor **good luck**, and in time-honoured tradition Verity Lambert leans forward and wishes the Director good luck  
the details. The rest is up to you. Au revoir, and **good luck!**" It is like this,' said the Town Clerk as they

---