


Corpus tools

How to make your writing more accurate

Aims


- ▶ To introduce two corpus tools:
 - www.lextutor.ca
 - IntelliText Leeds (<http://smlc09.leeds.ac.uk/itb/htdocs/Query.html>)
- ▶ To suggest ways in which these tools can help improve your writing
- ▶ To get first-hand experience in using these tools

What is a “corpus tool”?

- ▶ Corpus: a collection of texts
 - ▶ Concordance: collation of all occurrences of a word in a text or text collection
 - ▶ “Corpus tool” here: a website/interface/programme to create concordances or to extract other information from a corpus (e.g. collocations)
- 

Why use corpus tools?


- ▶ give you words in context
 - ▶ examples of grammatical structure
 - ▶ generally, more information than a dictionary

 - ▶ Example: Is “assert” followed by a *that*-clause or a noun / noun phrase?
- 

Tool 1: Lextutor

The screenshot displays the Lextutor website interface. At the top, a navigation bar contains the text "Compleat Lexical Tutor v.8.2 For data-driven language learning on the Web". Below this, a yellow banner features a search bar and the text "Lextutor". Two yellow arrows point to the "Lextutor" and "Lextutor" buttons in the top navigation bar. The main content area is divided into three columns: "LEARNERS", "RESEARCHERS", and "TEACHERS". The "LEARNERS" column lists tools like "Text", "Lext. Learn", "Group Lex", "Kil-Pending", "Context Grammar", and "Context Writer". The "RESEARCHERS" column lists tools like "Browse", "Vocabulary", "Comprehension", "RT Dumps", "RT Dumps", "Word Lists", and "Research Data". The "TEACHERS" column lists tools like "Text Tools", "Feedback", "Frequency", "Text Lex. Compare", "NativeGlossary", "New Words", and "Wordlists". The bottom of the page features the Lextutor logo and the text "© 2008-2010 Lextutor".

How it works...

- ▶ You type your query in the box at the top...
 - ▶ ...and click the red box to the far right.
 - ▶ You then get concordance lines:
- 

Concordance lines



The image shows a screenshot of a concordance search interface. At the top, there is a title bar and a menu bar. Below that, there is a search bar with the text "Concordance for equal SORT". A purple arrow points from the search bar down to the right column of the concordance results. The results are displayed in two columns, with the left column containing the search terms and the right column containing the corresponding text. The text is in a serif font and is arranged in a grid-like format. The search terms are in a smaller font size than the text they are searching for. The interface includes various controls such as search buttons, a list of search terms, and a list of results. The overall appearance is that of a classic desktop application from the late 1990s or early 2000s.

Which corpus?



Screenshot of a web browser displaying a document with a table of text. The browser's address bar shows a URL starting with 'http://www.assrt.com/'. The document content is a table with multiple columns of text, including a large purple arrow pointing downwards. The text in the table is partially obscured by the arrow and is mostly illegible due to blurring. The text appears to be a list of items or a table of contents, with some words like 'ASSRT' and 'ASSRT' visible. The browser's interface includes a search bar, navigation buttons, and a status bar at the bottom.

Which corpus?

The image shows a screenshot of a concordance search interface. At the top, the title bar reads "Concordance for equal ASSERT". Below the title bar, there are search filters and controls, including a search term "what I with left of key" and a search range "1-1000000". The main area displays a list of search results, each consisting of a line number, a search term, and a snippet of text. Two red arrows point to specific entries in the list, highlighting the search term "what I with left of key" and the corresponding text snippet.

Line	Search Term	Text Snippet
11	what I with left of key	... what I with left of key ...
12	what I with left of key	... what I with left of key ...
13	what I with left of key	... what I with left of key ...
14	what I with left of key	... what I with left of key ...
15	what I with left of key	... what I with left of key ...
16	what I with left of key	... what I with left of key ...
17	what I with left of key	... what I with left of key ...
18	what I with left of key	... what I with left of key ...
19	what I with left of key	... what I with left of key ...
20	what I with left of key	... what I with left of key ...
21	what I with left of key	... what I with left of key ...
22	what I with left of key	... what I with left of key ...
23	what I with left of key	... what I with left of key ...
24	what I with left of key	... what I with left of key ...
25	what I with left of key	... what I with left of key ...
26	what I with left of key	... what I with left of key ...
27	what I with left of key	... what I with left of key ...
28	what I with left of key	... what I with left of key ...
29	what I with left of key	... what I with left of key ...
30	what I with left of key	... what I with left of key ...
31	what I with left of key	... what I with left of key ...
32	what I with left of key	... what I with left of key ...
33	what I with left of key	... what I with left of key ...
34	what I with left of key	... what I with left of key ...
35	what I with left of key	... what I with left of key ...
36	what I with left of key	... what I with left of key ...
37	what I with left of key	... what I with left of key ...
38	what I with left of key	... what I with left of key ...
39	what I with left of key	... what I with left of key ...
40	what I with left of key	... what I with left of key ...
41	what I with left of key	... what I with left of key ...
42	what I with left of key	... what I with left of key ...
43	what I with left of key	... what I with left of key ...
44	what I with left of key	... what I with left of key ...
45	what I with left of key	... what I with left of key ...
46	what I with left of key	... what I with left of key ...
47	what I with left of key	... what I with left of key ...
48	what I with left of key	... what I with left of key ...
49	what I with left of key	... what I with left of key ...
50	what I with left of key	... what I with left of key ...

Change key

Key

equal

assert

in

BFC Commerce (3 Dec)

sorted

1 of 1

right

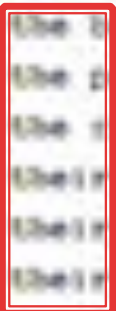
4 of 4

lized to 6 per million for comparison)

Click keyword for larger context

its faults. Groups who have normally felt able to [ASSERT](#) a strong bargaining position did
determinedly attempted to block Self's attempts to [ASSERT](#) control, which he saw as unaccept
rship in the performance of the business owned can [ASSERT](#) itself. If, however, the authori
democracy; not enough for the majority Report to [ASSERT](#) simply as a statement of belief,
be a very bold man, a Karl Marx indeed who would [ASSERT](#) that, for each and every woman a
set a subtle shift of emphasis. It is one thing to [ASSERT](#) that a consequence of sustained
de. Regulationist theories Regulationist theories [ASSERT](#) that advanced capitalist economi
is not the key marketing tool for books. And they [ASSERT](#) that discounting will only lead
chael. Although markets are treated globally, they [ASSERT](#) that overseas operations act sin
ly sells it to P. G is still the owner and he can [ASSERT](#) that ownership. In Central Newbu
ed in the previous section.) Finally, monetarists [ASSERT](#) that the economy (again in the s
constraints on their actions. It seems natural to [ASSERT](#) that the extent to which manager
le minimum. It is something altogether stronger to [ASSERT](#) that the unemployment rate can o
be theoretical underpinnings of their model. They [ASSERT](#) that the years selected for comp
case against markets often goes even further, to [ASSERT](#) the basic incompatibility of mar
y structure, managers are obliged continually to [ASSERT](#) the power of their position, whi
rights of labour, while trade unionism existed to [ASSERT](#) the rights of labour against the
the emergency was over and that they could safely [ASSERT](#) their "rights". Just so, Mr Boaz
re face too often and will hide their time to re - [ASSERT](#) their authority. You may ask whe
it is to be ineffective would enable them better to [ASSERT](#) their independence in future. Ge
ope and professions and when these groups seek to [ASSERT](#) themselves than the organisation
of the ...

007. ions methods. Regulationist theories Regulationist theories ASSERT that advanced capit
008. accounting is not the key marketing tool for books. And they ASSERT that discounting w
009. a truly global. Although markets are treated globally, they ASSERT that overseas oper
010. authorizedly sells it to F. O is still the owner and he can ASSERT that ownership. In
011. one outlined in the previous section.) Finally, monetarists ASSERT that the economy i
012. gets face constraints on their actions. It seems natural to ASSERT that the extent to
013. unavoidable minimum. It is something altogether stronger to ASSERT that the unemploye
014. sical of the theoretical underpinnings of their model. They ASSERT that the years wil
015. nally, the case against markets often goes even further, to ASSERT the basic incompe
016. n authority structures, managers are obliged continually to ASSERT the power of their
017. since the rights of labour, while trade unionism existed to ASSERT the rights of labo
018. eign that the emergency was over and that they could safely ASSERT their "rights". Ca
019. ord to lose face too often and will hide their true face to re - ASSERT their authority. T
020. en showing it to be ineffective would enable them better to ASSERT their independence
021. ferent groups and professions and when these groups seek to ASSERT themselves then th
022. at the end of the day will have it happen. While one cannot ASSERT unequivocally that
023. control, and is best understood as a counter - attempt to ASSERT workers' own contr
024. ou need to be selective and judge when it is appropriate to ASSERT your rights. Anoth



RIGHT IMMEDIATE 1 WD COLLOCS (Freq>=2) for assert (assoc=)


that=9 the=3 their=3



ALL POTENTIAL COLLOCS (n=7) for assert (assoc=) (freq>=4, content only, within 4 words either side)

none

Answer to example:

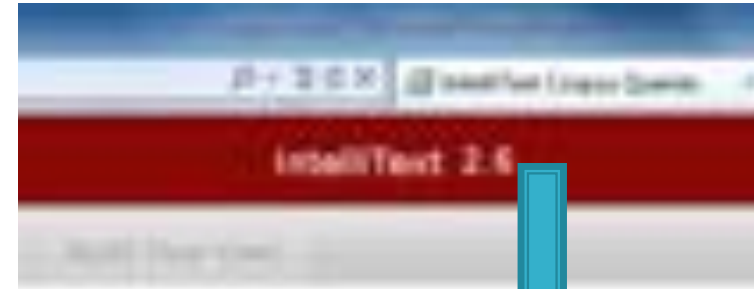
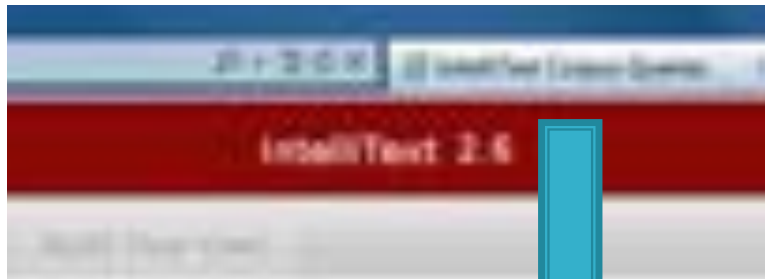
- ▶ “assert” can be used both with a *that*-clause (slide 11) and with a noun or noun phrase (slide 12).
 - ▶ To get an answer to the question “followed by...?”, remember to sort your concordance lines by the word following the key word, i.e. “right” (slide 11).
- 

- ▶ If you get “Zero hits”, try another corpus or another spelling. Or maybe the word doesn’t exist –> go to <http://dictionary.cambridge.org/> and type the word there, and hopefully they will suggest an alternative spelling...

Suggested uses: Lextutor

- ▶ Generally, to check your intuition (i.e. if you think you know already)
- ▶ To check whether a word is really used in a context that you think it is used in
- ▶ To check what is used with a given verb:
 - which preposition
 - Gerund (-ing) or *that*-clause
 - ...
- ▶ Main advantage: Quick and easy access

Tool 2: IntelliText



Search the Standard Corpora	0-0
Build or Search Your Own Corpora	0-0
Welcome	0-0
Acknowledgements	0-0

Arabic	0-0
Chinese	0-0
English	0-0
French	0-0
German	0-0
Italian	0-0
Japanese	0-0
Kannada	0-0
Portuguese	0-0
Russian	0-0
Spanish	0-0
Ukrainian	0-0
Parallel (Translation) Corpora	0-0

Different corpora – which one to use?

IntelliText 2.6

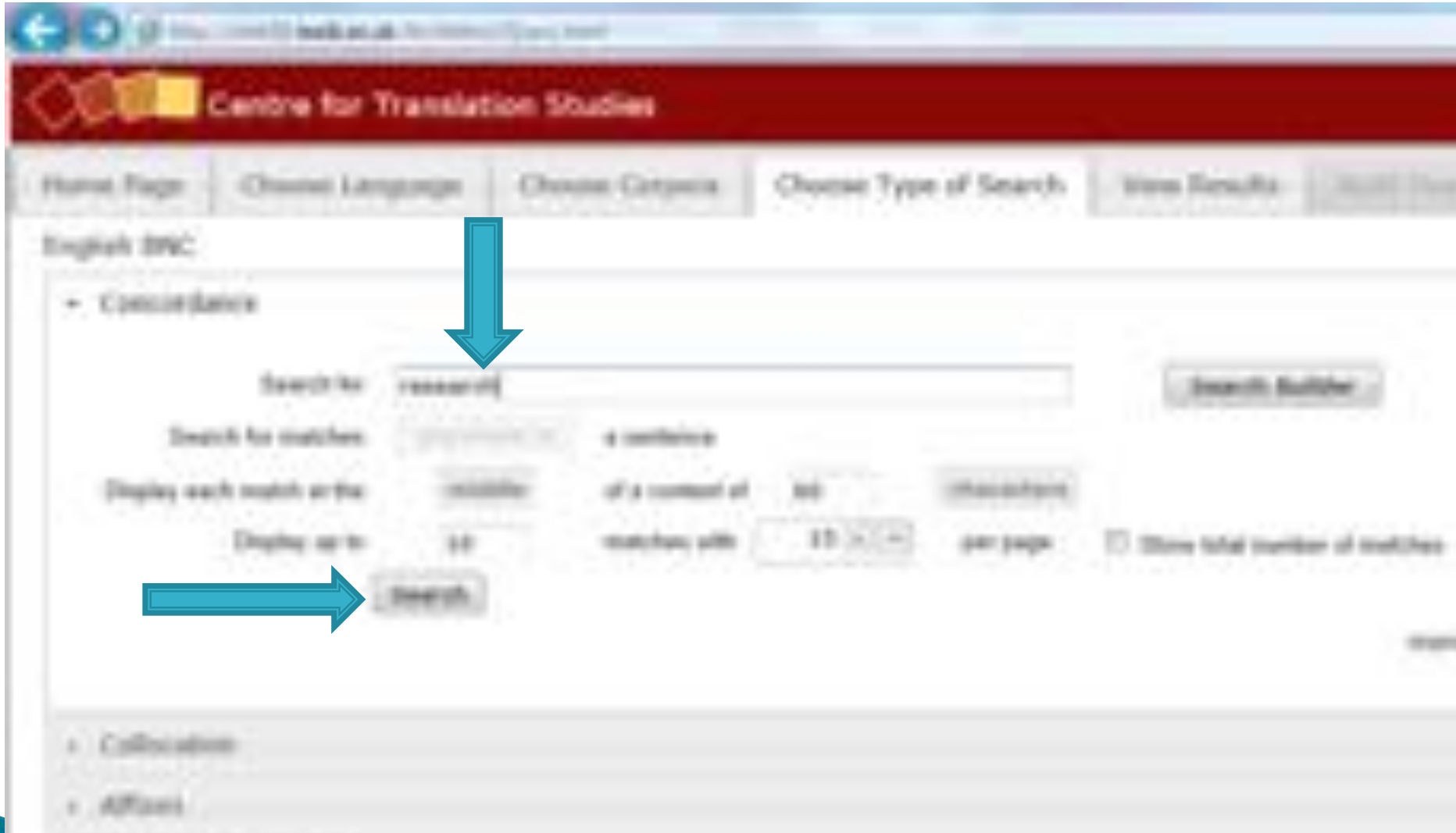
Search Type of Search

Corpus	Description	Size (Millions of Words)	Link
ERIC	A balanced corpus of British English up to 1994	100.0	http://www.ericp.org.uk/
INTERNET-EN	Internet texts in English	150.0	http://www.intellitext.com/uk/ukfichsearch.pdf
URONAC	UR based web pages	200.0	http://www.uranet.com/uk/ukfichsearch.pdf
WIKI-EN	Wikipedia entries in English	943	http://www.intellitext.com/uk/ukfichsearch.pdf
REUTERS	Collection from newswires	90.0	http://www.intellitext.com/uk/ukfichsearch.pdf
NEWS-GB	Collected from British newspapers	200.0	http://www.intellitext.com/uk/ukfichsearch.pdf
BLOOM-EN	Fictional texts	500.0	http://www.intellitext.com/uk/ukfichsearch.pdf
BROWN	A corpus of American English up to 1961	1.0	http://www.intellitext.com/uk/ukfichsearch.pdf
BE00	British corpus from 2000	1.1	http://corpus.lancs.ac.uk/2000/
AME00	American English corpus from 2000	1.1	http://corpus.lancs.ac.uk/2000/
BREAST-EN	Breast cancer corpus	0.9	
DRUGS-EN	Drug addiction corpus	0.9	

Next >



Type your search query



The screenshot shows the search interface of the Centre for Translation Studies. At the top, there is a navigation bar with buttons for "Home Page", "Choose Language", "Choose Corpus", "Choose Type of Search", "View Results", and "Print Page". Below this, the text "English BNC" is displayed. A section titled "Concordance" contains a search form. The "Search for" field contains the text "research". To the right of this field is a "Search Button". Below the search field, there are several options: "Search for matches" (set to "anywhere"), "Display each match on the" (set to "middle"), "of a context of" (set to "40"), "Display with" (set to "40"), "matches with" (set to "10"), "per page" (set to "40"), and a checkbox for "Show total number of matches". A "Search" button is located at the bottom left of the form. Two blue arrows are overlaid on the image: one pointing down to the search input field and another pointing right to the "Search" button.

Centre for Translation Studies

Home Page | Choose Language | Choose Corpus | Choose Type of Search | View Results | Print Page

English BNC

Concordance

Search for: Search Button

Search for matches: anywhere 4 matches

Display each match on the: middle of a context of: 40

Display with: 40 matches with: 10 per page: 40 Show total number of matches

Search

Here are your concordance lines.


Centre for Translation Studies

Home Page | Choose Language | Choose Category | Choose Type of Search | View Results

Concordance for search:

Word	left	right
20	announced to the press that " a	visit to Sri Lanka would be considered
20	an exhilarating preparation for	degrees in Law, Philosophy and
20	very life blood of Amnesty is its	teams — there are currently
20	were of any team or even	experts. In those days there was
20	There was a small core of paid	staff who were responsible for
20	as well "volunteers worked on	throughout their past counterparts
20	generated by the many requests and	like Amnesty sends to countries
20	assessment in carrying out their	. As it relates to Argentina in
20	background" " When you've done	on a country for years, you also
20	subsequently announced that " a	visit to Sri Lanka would be considered

Search builder

- ▶ Choose:
 - Part of speech (verb? noun? preposition?)
 - For verbs: which verb form?
 - Intermediates (how many words in between the two words you are searching for?)
 - Base form of the word or this exact form?
 - ▶ This results in a very precise search, and it can give you exactly the word you needed.
- 

Search builder

The screenshot shows a web browser window with the address bar displaying a URL. The page header is a dark red bar with the text "Centre for Translation Studies" and a logo of four colored diamonds. Below the header is a navigation menu with buttons for "Home Page", "Choose Language", "Choose Corpus", "Choose Type of Search", and "View Results". The main content area is titled "English BNC" and contains a search form. The form has a "Search for:" field with the text "research", a "Search Builder" button, and several other options: "Search for matches" (set to "anywhere"), "Display each match as the" (set to "middle"), "of a context of" (set to "50"), "Display as" (set to "10"), "matches with" (set to "10"), "per page" (set to "10"), and a checkbox for "Show total number of matches". A large blue arrow points down from the "View Results" button in the navigation menu to the "Search Builder" button in the search form. At the bottom of the page, there are links for "Collaboration" and "Affairs".

Search Builder

<input type="text"/>	<input checked="" type="checkbox"/> search for word in this form <input type="checkbox"/> search for all forms of this base word	<input type="button" value="Full Edit"/> <input type="button" value="Clear Full"/>
<input type="button" value="Add Intermediate"/>	<input type="button" value="Add Another Word"/>	
<input type="text"/>	<input checked="" type="checkbox"/> search for word in this form <input type="checkbox"/> search for all forms of this base word	<input type="button" value="Full Edit"/> <input type="button" value="Clear Full"/>
<input type="button" value="Add Intermediate"/>	<input type="button" value="Add Another Word"/>	
<input type="text"/>	<input checked="" type="checkbox"/> search for word in this form <input type="checkbox"/> search for all forms of this base word	<input type="button" value="Full Edit"/> <input type="button" value="Clear Full"/>
<input type="button" value="Add Intermediate"/>	<input type="button" value="Add Another Word"/>	

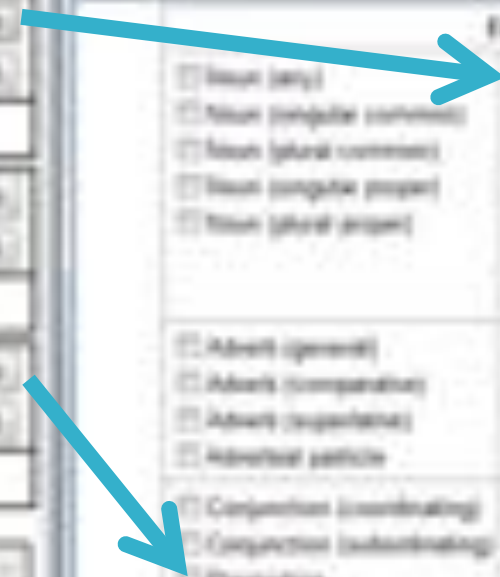
OK Apply Clear Cancel Doing

English Morphological Features - Windows Internet Explorer

http://www.ellie.com/uk/uk/ellie/eng/words/feature/2002/000011

English Parts of Speech		
<input type="checkbox"/> Noun (any)	<input type="checkbox"/> Verb (any part)	<input type="checkbox"/> Adjective (general)
<input type="checkbox"/> Noun (single common)	<input type="checkbox"/> Verb (base)	<input type="checkbox"/> Adjective (comparative)
<input type="checkbox"/> Noun (plural common)	<input type="checkbox"/> Verb (past tense)	<input type="checkbox"/> Adjective (superlative)
<input type="checkbox"/> Noun (single proper)	<input type="checkbox"/> Verb (ing)	
<input type="checkbox"/> Noun (plural proper)	<input type="checkbox"/> Verb (past participle)	
	<input type="checkbox"/> Verb (ed)	
	<input type="checkbox"/> Modal auxiliary	
<input type="checkbox"/> Adverb (general)	<input type="checkbox"/> Pronoun (relative)	<input type="checkbox"/> Wh (general)
<input type="checkbox"/> Adverb (comparative)	<input type="checkbox"/> Pronoun (personal)	<input type="checkbox"/> Wh (possessive pronoun)
<input type="checkbox"/> Adverb (superlative)	<input type="checkbox"/> Pronoun (possessive)	<input type="checkbox"/> Wh (interrogative)
<input type="checkbox"/> Adverbial particle		<input type="checkbox"/> Wh (adverb)
<input type="checkbox"/> Conjunction (coordinating)	<input type="checkbox"/> Determiner (general)	<input type="checkbox"/> Possessive particle
<input type="checkbox"/> Conjunction (subordinating)	<input type="checkbox"/> Determiner (specific)	<input type="checkbox"/> Foreign word
<input type="checkbox"/> Preposition	<input type="checkbox"/> It's (possessive marker)	<input type="checkbox"/> Quot (quote marker)
<input type="checkbox"/> Interjection	<input type="checkbox"/> Cardinal number	<input type="checkbox"/> Symbol
	<input type="checkbox"/> Ordinal number	<input type="checkbox"/> End of sentence

OK Cancel Doing



More interesting concordance lines

Centre for Translation Studies

Home Page | Choose Language | Choose Corpus | Choose Type of Search | View Results

Concordance for [your query] [with search type="KW"]

Word	left	match	right
820	background. "When you're	done research on	a country for years, you also
825	separated by cognitive barriers while	publishing research in	what was ultimately the main source
828	director has presented a paper today	cutting research into	the Police and Criminal Evidence
828	agreement in the case of sociologists	conducting research on	the police and long for a return
828	though to the college though, to	do research on	their own account. The Staff College
828	is." to ask whether anyone is	doing research on	the police, what degree of co-operation
828	they have met ... and to	encourage research by	the police themselves." Once again
828	of though, collect, collate and	introduce research of	interest to the police service
831	topic in the (Pro-) its argument. To	begin research on	a fertilised egg unless it is
833	department's internationally	disseminated research in	comparative endocrinology, analysing

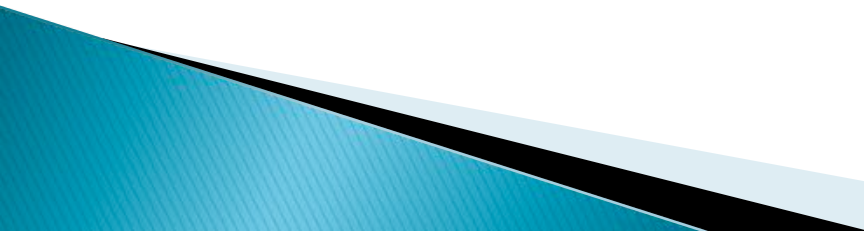
Collocations

The screenshot shows the website interface for the Centre for Translation Studies. At the top, there is a navigation bar with options: Home Page, Choose Language, Choose Corpus, Choose Type of Search, and View Results. Below this, the page is titled "English BNC" and "Concordance". A search form is visible with the following fields and options:

- Search for:
- Search for matches: a sentence
- Display each match with: of a context of characters
- Display with: matches with per page
- Show total number of matches

A large blue arrow points downwards from the search form area towards the "Collocation" and "Affix" options at the bottom of the page.

Collocations

- ▶ To find which words go together, e.g. **what do you actually “do” with “research”?** Which is the word most frequently used together with research (and therefore probably, but not necessarily, the one you need)?
 - ▶ Which word do you need a collocation for?
 - ▶ Does it come before or after the word you have?
 - ▶ Which kind of word do you need for your word? → “Collocate Part of Speech”
- 

Build your search

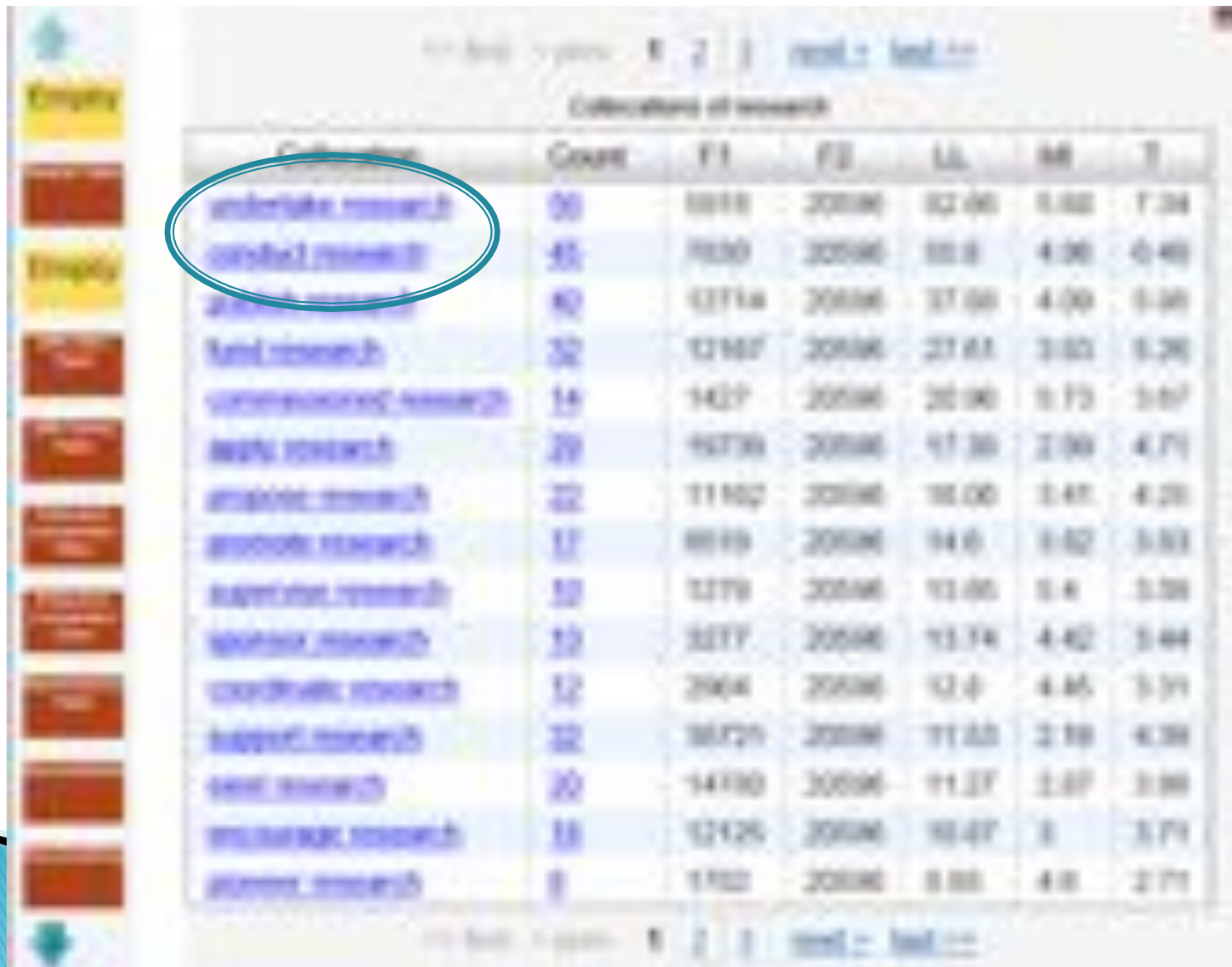
The screenshot shows the search interface of the Centre for Translation Studies. At the top, there is a red header with the logo and the text "Centre for Translation Studies". Below the header, there are navigation tabs: "Home Page", "Choose Language", "Choose Corpora", "Choose Type of Search", and "View Results". The "Choose Type of Search" tab is selected, and the "English BNC" corpus is chosen. The search type is set to "Collocation".

Annotations with blue arrows and a circle highlight key search options:

- A downward arrow points to the search input field containing the word "research".
- A diagonal arrow points to the search input field.
- A circle highlights the "words to the left of" dropdown menu.
- An upward arrow points to the "Collocate Part of Speech" dropdown menu, which is set to "v*".
- Another upward arrow points to the "Find Collocates" button.

Other visible elements include a "Search" button, a "Search Results" button, and a "Words to the right of" dropdown menu.


Here are your collocations




Collocations of research

Collocation	Count	F1	F2	L1	M	T
understand research	58	1818	2008	32.85	5.82	7.24
conduct research	45	1630	2008	33.8	4.36	6.49
do research	40	13714	2008	37.88	4.08	6.38
find research	32	12187	2008	27.81	3.82	5.28
conducting research	19	1427	2008	28.88	8.73	3.87
study research	29	15738	2008	47.38	2.88	4.71
analyze research	22	11182	2008	38.88	3.47	4.28
provide research	17	8118	2008	34.8	1.82	3.81
experiment research	12	1278	2008	33.88	8.4	3.28
perform research	12	1217	2008	33.74	4.42	3.44
conducting research	12	284	2008	32.8	4.45	3.31
gather research	12	18721	2008	71.82	2.18	4.38
test research	20	14780	2008	41.27	2.87	3.88
investigate research	18	12125	2008	38.47	3	3.71
perform research	8	1782	2008	8.88	4.8	2.71

Suggested uses: IntelliText

- ▶ Generally, to **get answers if you don't know**
 - ▶ To **find out** about collocations (words that go together)
 - ▶ And of course, this is just the start. I'm sure you will find this useful; if after a while you want to learn more about the possibilities, please ask me!
- 

Need more context?

- ▶ In both interfaces, you can view the word in the wider context, to help you decide whether it is used like you think it is used (sometimes half a sentence is not enough to see this).
 - ▶ Lextutor: click on the keyword in the line you want to see in context
 - ▶ IntelliText: click on the text number to the left of the concordance line
- 

Summary

- ▶ Lextutor:
 - Quick and easy
 - For single-word queries
 - To **confirm** immediate collocations
- ▶ IntelliText:
 - Takes a bit of clicking to access the corpus
 - For more **sophisticated** queries
 - Two-word queries; non-immediate collocations; to check prepositions
 - Access to the full BNC (British National Corpus) -> much broader basis for your search