Using a Corpus as a Resource for Writing

Information for Students

Maggie Charles
maggie.charles@lang.ox.ac.uk

1 Introduction: Corpus and Concordance

The purpose of this material is to introduce you to working with a corpus and software that can help you with your writing.

Here are some questions you may sometimes ask.

- *Can I say …? Is this phrase often used?*
- *What preposition should I use?*
- *Should I use 'a' or 'the' or neither?*
- *How do I discuss the literature?*
- *How do researchers present their results?*

Using a corpus provides information to help you answer questions like these.

What is a corpus?

A corpus (plural: *corpora*) is a collection of electronic texts. Corpora are built with a specific purpose in mind and are designed according to specific criteria. For example, if you want to check the language used in your field, you can build a corpus of relevant research articles written by experts. Corpora are accessed by using text analysis software.

What does the software do?

Text analysis software analyses the texts in the corpus automatically and presents the results on screen. It provides several different tools for examining the data, the most widely-used being the **concordancer**.

What does the concordancer do?

The concordancer searches the corpus for every instance of a word or phrase you specify and presents each one with its context in a line on screen.

The **search term** appears in the centre, usually with about 5 or 6 words either side of it. This is called a 'keyword in context' or **kwic** concordance.

Student Question

- *Can I say 'a broad variety of…'*?

Example 1 gives part of a concordance on the search term *variety*, retrieved from a 500,000 word corpus of Oxford theses. The concordance shows that the most frequently used phrase is '*a **wide** variety of…*'. The adjective *broad* does not occur at all with the noun *variety* in this corpus. Words often occur in combinations which are more or less fixed, that is they **collocate** with other words. Using a corpus and concordance can help you find information on these collocations, which is often not available in a dictionary or reference grammar.

Example 1: Concordance on the Word 'v*ariety*'

1	s that the carbon black in one batch of a	**given variety**	(e.g. Raven 430) has a large
2	strong 1993: 13-15). <ptr3> Although the	**historical variety**	of ideas of international
3	e neatness of a cycle imposed by Vico on the	**huge variety**	of human history - then we are
4	the severity. With these qualifications a	**large variety**	of samples have been measured
5	ommensurable and all contributing to the	**rich variety**	that makes up Humanity. Mazzini
6	s</head> These materials exhibit the	**richest variety**	of chemistry including accept
7	arkises wicked by-blow?"'(GR 171. Here the	**standard variety**	of English at the beginning
8	s demonstrated that it depends upon a	**wide variety**	of factors. This has enabled the
9	ools for future studies of treatments on a	**wide variety**	of materials. The details of the
10	ll ate widely differing diets. Yet with a	**wide variety**	of foods available to north west
11	the samples measured would have resulted in a	**wide variety**	of decay forms and decay rates
12	the biosynthesis of amino acids occurs via a	**wide variety**	of metabolic routes, the carbon
13	e of taking several thousand STM images of a	**wide variety**	of metallic surfaces [3.15].
14	asured at raised temperature (160(C), from a	**wide variety**	of natural samples, could be 23
15	43 0.5. As shown in Figures 7.4 and 7.5, a	**wide variety**	of shine-plot forms are possible
16	pe was used<ptr3>. <p>Because of the	**wide variety**	of electron microscopes availabl
17	oduces gram quantities of material, the	**wide variety**	of conditions within the confine
18	ining the dose response characteristics of a	**wider variety**	of samples; (vii) the
19	wide range of resolved emission bands and a	**wider variety**	of storage times/temperatures.
20	of n (to perhaps exponential or Gaussian) a	**wider variety**	of decay forms could be

Reading a concordance

You may have noticed that reading a concordance is different from normal reading. You read down the page, not across and you look at the words to the right and left of your search word. It is not necessary to read every word.

TIP

When you read concordances, look for **repeated regularities**.

They show you the patterns of language that are frequently used in your corpus.

How can a corpus help me?

- You can see many examples of a word or phrase at the same time.
- You can see which words and phrases are commonly used in your field.
- You can find information which may not be given in a dictionary or reference grammar.
- You can compare your writing with expert writing.
- You can deal with language problems in your own writing.

2 Building your own Personal Corpus

There are two types of corpora you can build that may be useful for your writing.

1. **A corpus of expert texts**, e.g. research articles in your field
    - Useful because the language is specific to your field and relevant to your own writing needs
2. **A corpus of your own writing,** e.g. chapters of your thesis or your research papers
    - Useful because you can examine your own writing automatically to help identify problems

What should I put into my corpus?

If you want to build a corpus of expert texts, select articles that are well-written and well-regarded in your field and choose a range of different writers.

When choosing files for your corpus, take account of any copyright issues.

Keep each text in a separate file and give it a file name that you can recognise easily e.g. Author_date (Charles_2013).

If you want to build a corpus of your own work, put each chapter of your thesis or each research article in a separate file and name them accordingly e.g. Chap_1.

TIP
Use short file names so that they don't take up too much screen space.

How many files do I need for an expert corpus?

In general, the larger the corpus, the better. This is because you will get more hits with a larger corpus. For example, if you see 100 hits of '*the increase in*', you can be more confident that '*in*' is a good choice of preposition than if you only see 5 hits. It's good to aim for about 30-50 research articles, but you can begin to use your corpus as soon as you have prepared some files.

How do I build my corpus?

To build a corpus, you need to convert the files you want to include into **plain text format**, because concordance software does not work with pdf, Word or HTML files.

You can convert files that are in **Word format** directly to plain text.

If you want to include files that are in **pdf format**, you may not get good results by converting them directly. In this case, convert the files first to Word format and then to plain text format.

Table 1 shows the steps for converting files.

Table 1: Three ways to convert files

Converting from Word to Plain Text File	**Converting from pdf or HTML to Plain Text File**	**Converting from pdf or HTML to Word**
1. Open the file you want to convert. 2. On the file menu, choose **Save As…** 3. At the bottom of the box click on **Save as Type**. 4. Scroll down and select **Plain Text** (.txt), then **Save.** 5. When the file conversion window appears, click **OK**.	1. Open the file you want to convert. 2. On the file menu, choose **Save As…** 3. Under **Save as Type**, select **Plain Text** (.txt), then **Save.**	1. Open the free on-line tool: http://www.convertfiles.com/ 2. Upload the file you want to convert. Choose output MS Word. Press convert. 3. After conversion, open the file and convert from Word to Text, as in Column 1.

When you have converted your files, create a separate folder for your corpus and save the text files there.

TIP

Check that each file has converted properly.

If not, convert to Word and then to text, or choose a different file.

Do I have to make any other changes to the files?

When you check your text files, you will see that they contain words that are not part of the running text (e.g. references, title). Graphic elements (e.g. tables, figures) may also look strange.

You can clean these elements from the corpus files, simply by deleting them manually, but it takes time and many student users do not find it worthwhile.

You can also part-clean the files by deleting just the title and author details, references and graphics.

TIP
Try out your 'dirty' corpus first and if you don't like it, clean it gradually later.

3 Using the Concordance Software *AntConc*

We will use the *AntConc* software, version 3.2.4, which is freely downloadable from

http://www.antlab.sci.waseda.ac.jp/antconc_index.html

There are versions for Windows, Mac and Linux. Further information is available on the website.

Download the software and save it onto your computer. You can put an icon on your desktop.

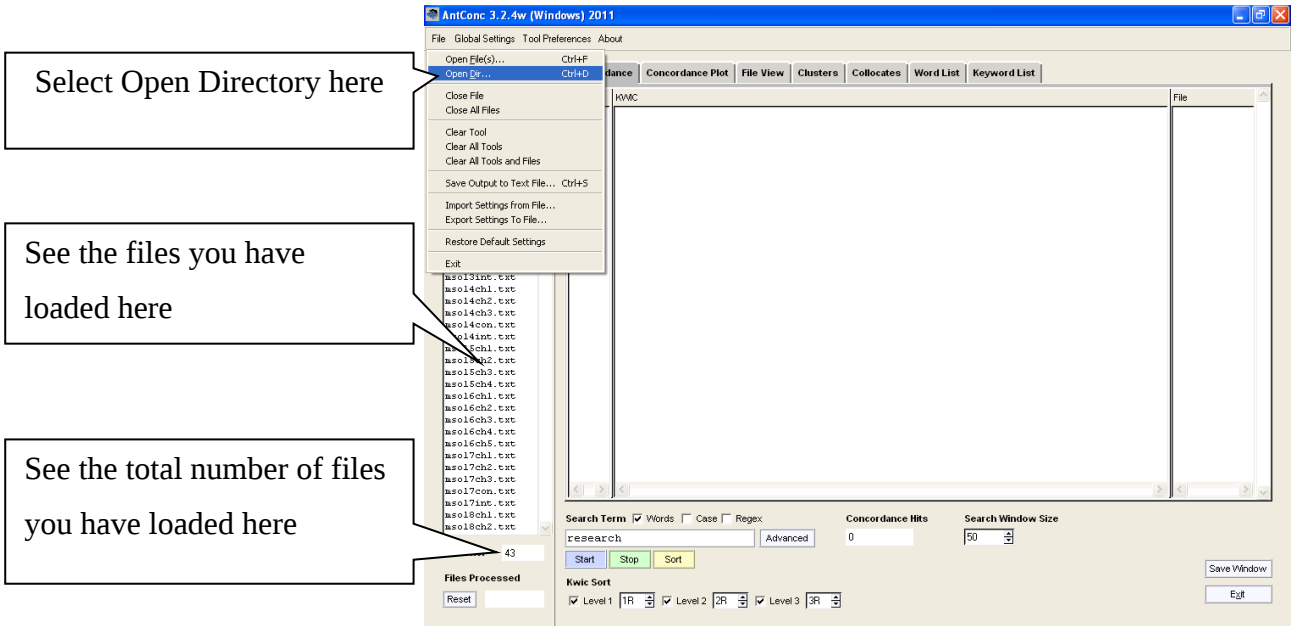
TIP
You can store AntConc and your corpus on the cloud (e.g. Dropbox) for access on-line.

3.1 Loading your Corpus

Before you can use *AntConc*, you have to load your corpus into the software.

1. On the **file menu**, select **open directory**.
2. Select the folder where you have put your corpus and click OK.
3. This loads all the files in your corpus folder.
4. To load only certain files, select **open file(s)**.

Screen Shot 1: *AntConc* Loaded with a Corpus



4 Using the Concordancer

When you have loaded your corpus, you are ready to use the concordancer.

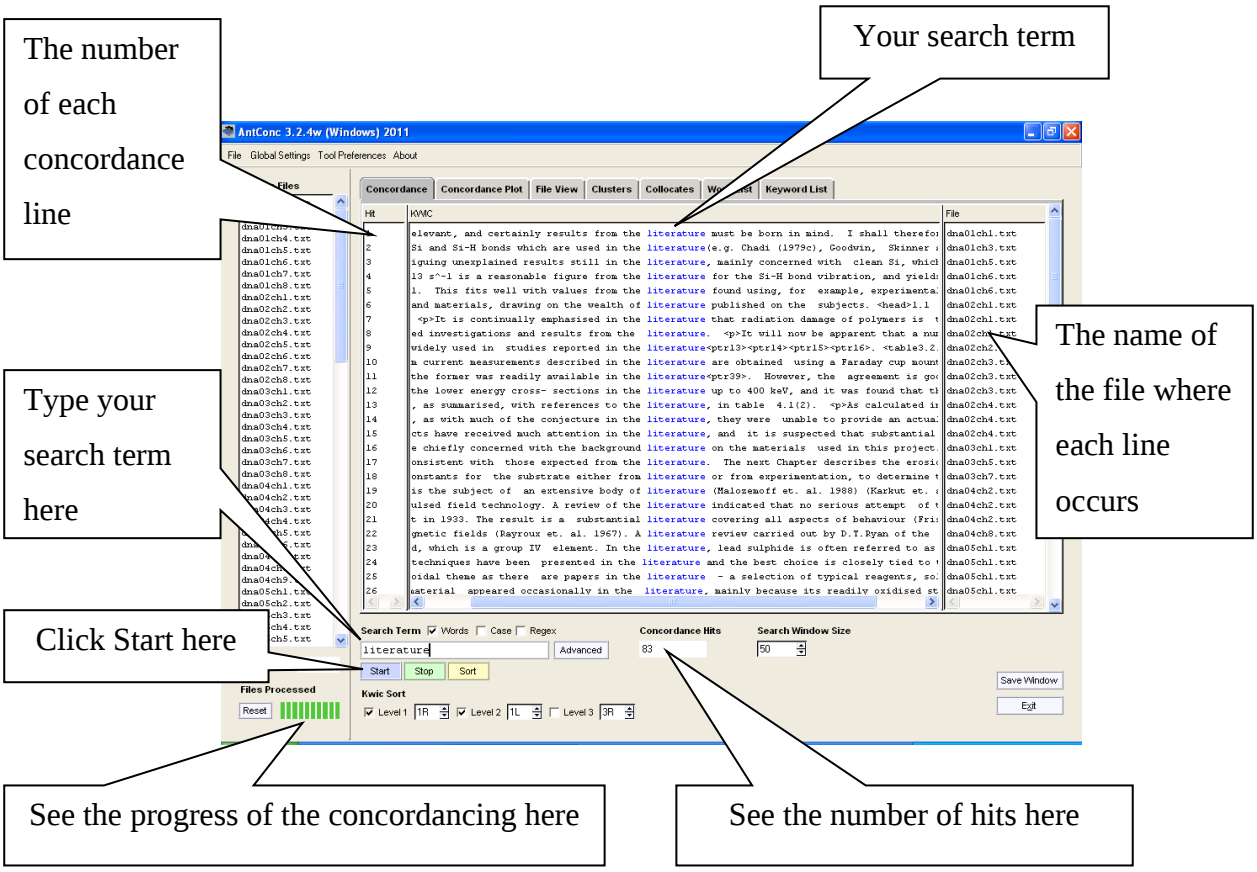
4.1 Making a Concordance

1. Type the word or phrase you want to find in the **Search Term** box.
2. Click the blue **Start** button below the box.
3. The **concordance lines** appear in the main window, with your search term highlighted in the centre.

Student Questions

- *How do I use the noun 'literature'?*
- *What preposition should I use after the noun?*
- *Should I use 'a', 'the' or neither before the noun?*

Screen Shot 2: Concordance on the Word '*literature*'



- **Files Processed** shows the progress of the concordancing with green bars.
- The **Concordance Hits** box flashes blue with the word FINISHED when the software has finished concordancing. Then it shows you the number of times your search term occurs.
- If there are no hits, the **Concordance Hits** box flashes red with the words NO HITS and the window does not change.

TIP

If you don't get any hits and the search term is quite common, check your typing.

Although the concordance above does give you answers to the student questions, you may have found it quite difficult to find the relevant information. To see the information more clearly, it's best to sort the concordance lines so that hits of the same type are grouped together.

4.2 Sorting Concordance Lines

Concordances are particularly useful for showing the patterns of vocabulary and grammar that are associated with your search term. You can see these clearly when you sort the concordance lines. You can sort the lines alphabetically by one, two or three words. The first sort is called Level 1, the second, Level 2 and the third, Level 3. For example, to see the preposition used after the noun '*literature*', we sort by the first word on the right. This is a Level 1 sort on 1R. To see whether to use '*a*', '*the*' or neither, we sort by the first word on the left of the noun, so we also make a Level 2 sort on 1L.

1. Sort the concordance lines by the words to the right or left of your search term by using the **Kwic Sort** boxes below the Start button. Choose the sort position you want by clicking the up or down arrows on the right of each Level box.

   0: sorts by your search term.

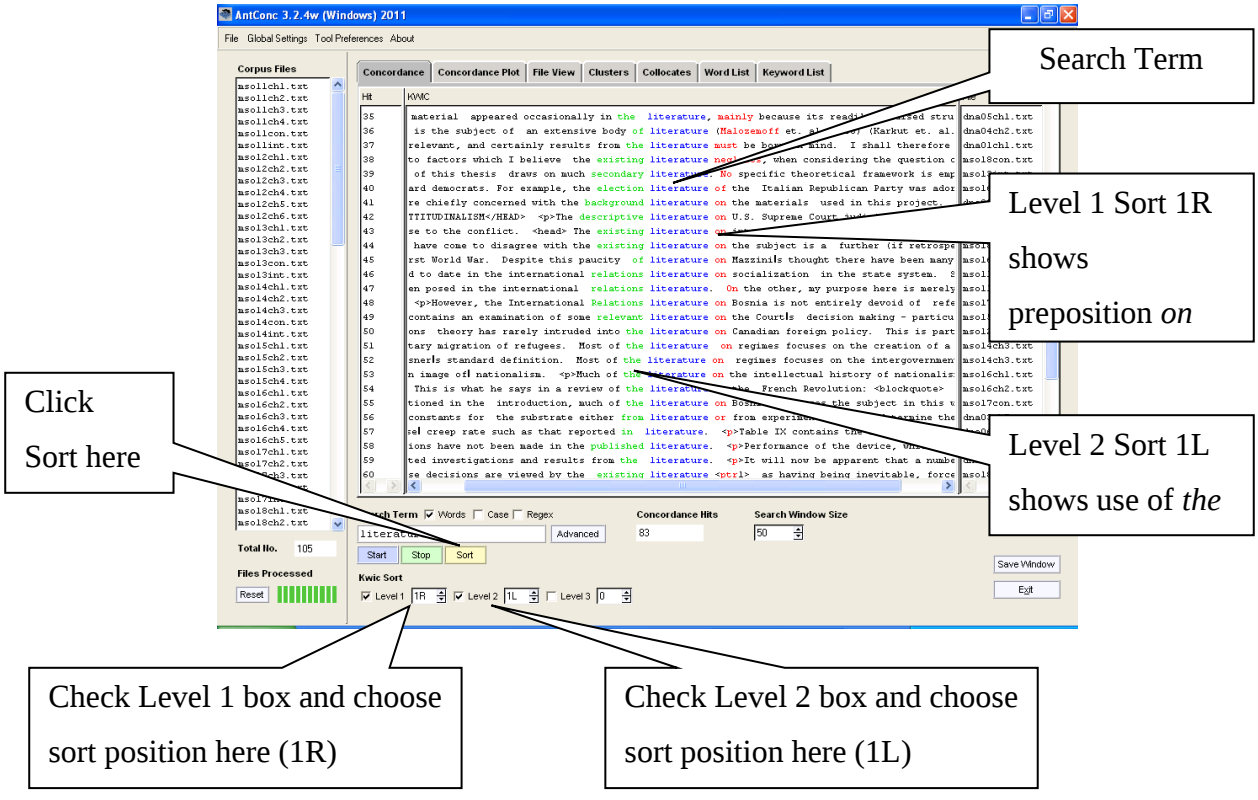
   1R: sorts by the first word on the right of the search term

   2R: sorts by the second word on the right of the search term and so on.

   1L: sorts by the first word on the left of the search term and so on.

2. To make Level 2 or Level 3 sorts, check the appropriate box and choose the sort position.
3. Click the yellow **Sort** button above the Kwic Sort boxes.
4. When the lines are sorted, the sort positions appear in different colours.

Screen Shot 3: Concordance on the Word '*literature*' Sorted by Level 1: 1R and Level 2: 1L



4.3 Viewing the Original File

If you want to see the wider context in which a certain example of your search term occurs, you can view the original file from which a specific concordance line is taken. This can be useful when you are concerned with broader issues about the way writing develops, for example, how an argument is constructed or how a writer organises a whole section of the text. From the concordance on '*literature*', you can move to an individual example, which shows you in more detail how that writer discusses others' research.

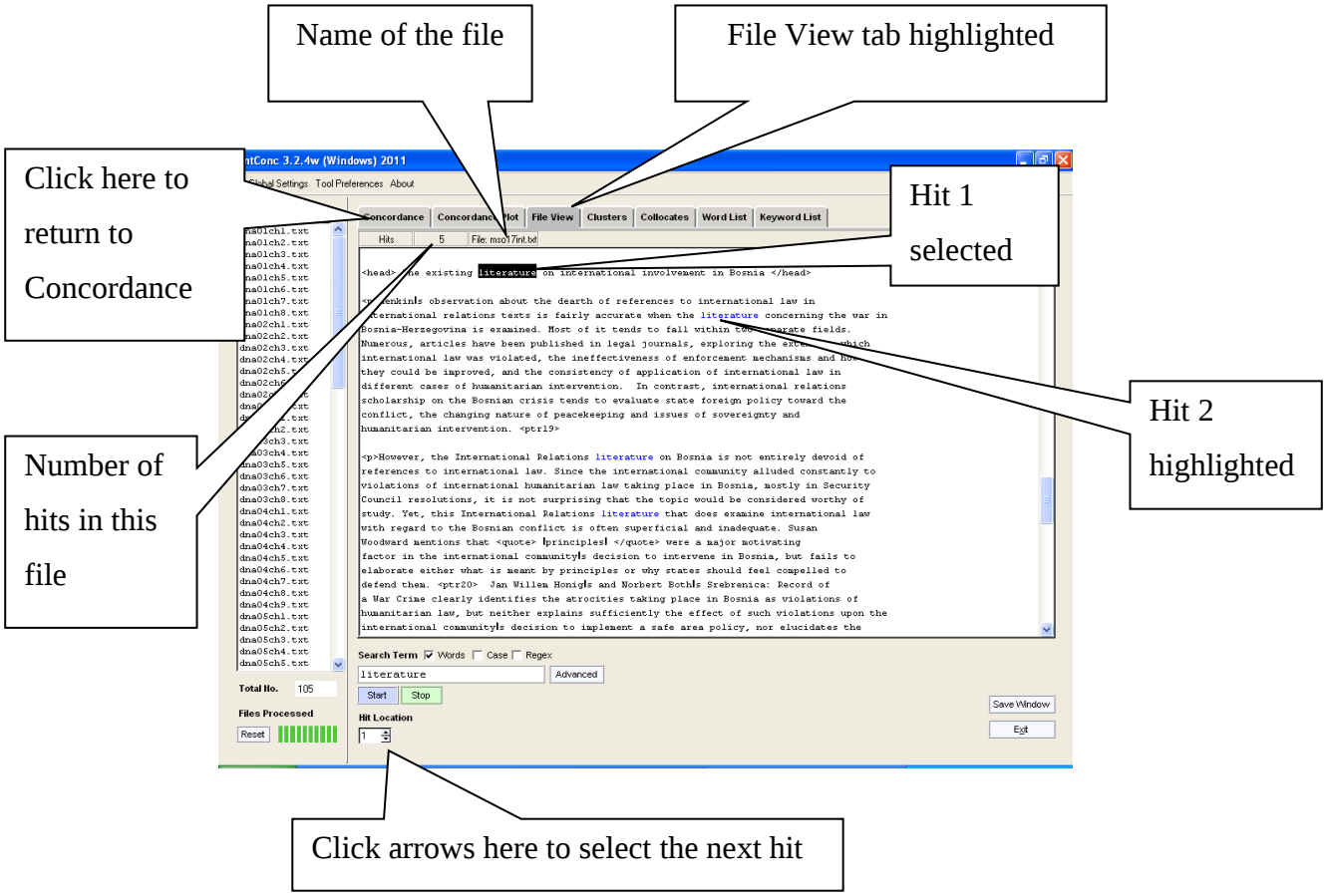
1. Move the cursor over the search term (here: '*literature*') in any one line.
2. When the cursor becomes a **hand**, click the search term. You will see the original file with the search term selected.
3. This is called **File View**. You will see that the **File View** tab at the top of the window is now highlighted.
4. The bar at the top of the window gives the total number of occurrences of your search term within that file and the name of the file.
5. To view each hit in that file, select the appropriate number by clicking the up and down arrows in the **Hit Location** box below the start and stop buttons on the bottom left.

TIP
To return to the concordance, click the **Concordance** tab at the top of the window.

Student Question

- *How do writers in my field deal with the literature?*

Screen Shot 4: File View of the Word *'literature'*



Viewing this original file shows you how the writer first presents an analysis of the literature as a whole (e.g. '*tends to fall within two fields*') and then critically evaluates one part of the research as '*superficial and inadequate*', supporting this assessment with references (e.g. *Woodward*) and critiques of specific work (e.g. *'…fails to elaborate…'*).

4.4 Making a Case Sensitive Search

You can search for a term in upper case (with capital letters) or lower case (no capitals). By default, *AntConc* retrieves both upper case and lower case forms. A case sensitive search is useful if you want to retrieve names, or words that begin a sentence.

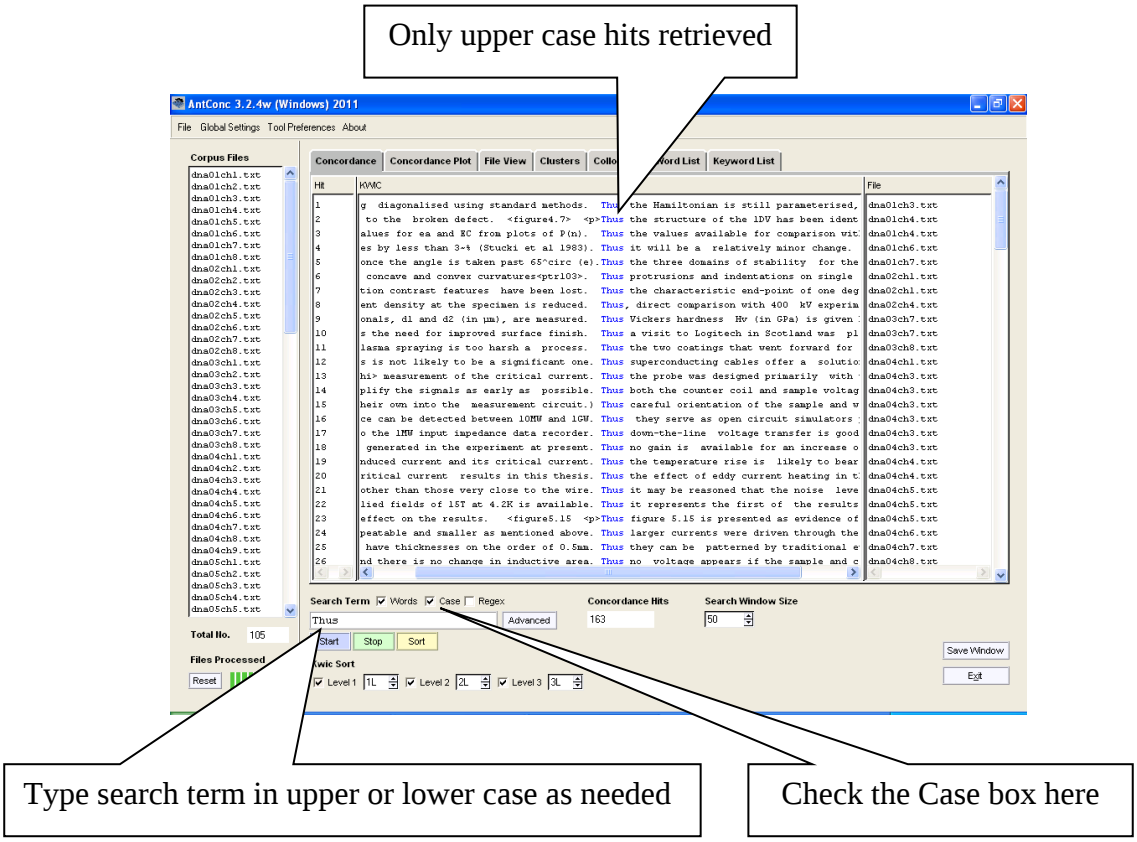
1. **Check** the **Case** box next to the Search Term Words box.
2. Type the search term with upper or lower case, as needed.
3. Click **Start**.

> **TIP**
>
> Remember to uncheck the Case box to return to normal searching.

Student Question

- *Can I begin a sentence with 'thus'?*

Screen Shot 5: Case Sensitive Concordance on the Word '*Thus*' in Upper Case



The concordance shows that '*thus*' can be used at the beginning of a sentence and provides many examples.

4.5 Making a Wild Card Search

A wild card is a symbol that represents any word or part of a word. One useful wild card is the star: *. When you use this symbol in your search, the software gives you any word or part of a word that occurs in that position. A wild card search can show all the different forms of a word, or it can retrieve all the words which occur in a specific position in a phrase you search.

Examples

1. The search term *conclu** retrieves all words beginning 'conclu': e.g. *conclude, concludes, concluded, conclusion, conclusions, conclusive, conclusively*. This type of search is useful to find word families sharing a root form and an aspect of meaning.
2. The search term **able* retrieves all words ending in 'able': e.g. *able, unable, capable, valuable*. This type of search is useful to find word classes like adjectives, adverbs or nouns.
3. The search term *it * necessary* retrieves e.g. *it is necessary, it was necessary, it seems necessary*. This type of search is useful to find the words that occur in the pattern searched.
4. The search term *it * * necessary* will retrieve e.g. *it is first necessary, it may be necessary*. Use more than one star to search for more than one word in a specific position.

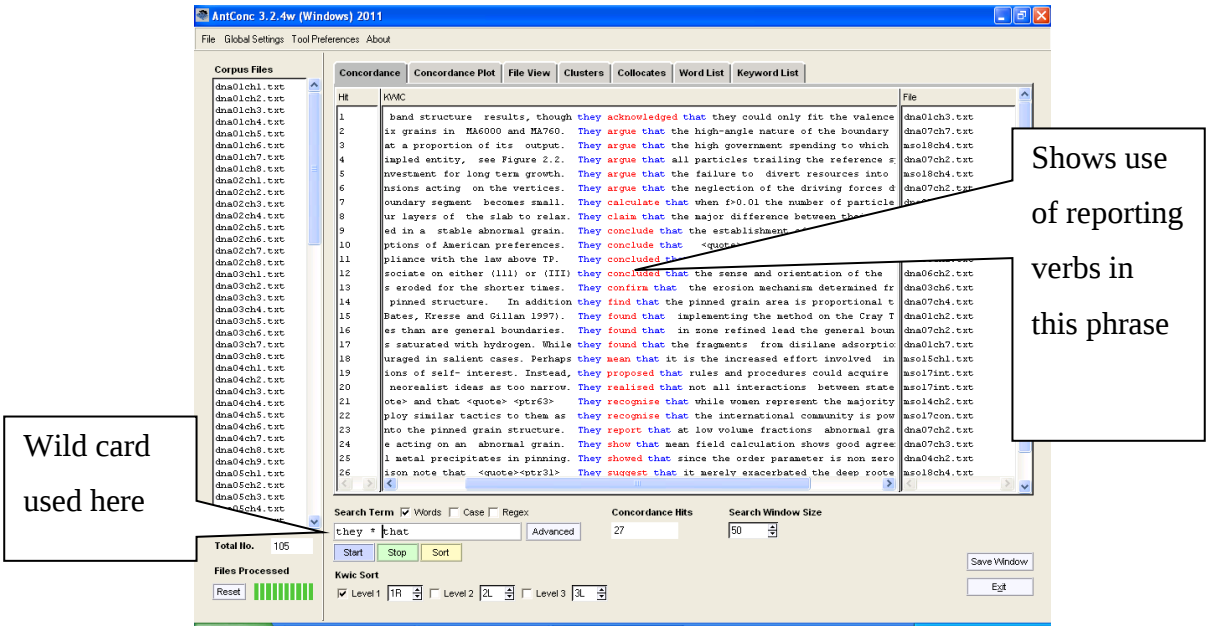
TIP

To retrieve a word in a specific position, put **a space on both sides** of the star.

Student Question

- *Which verbs can I use to report the research of other writers?*

Screen Shot 6: Wild Card Search on the Phrase *they * that*



The wild card search shows several different reporting verbs, including *argue*, *conclude*, *find* and *show*.

4.6 Making a Search for Two or More Terms at the Same Time

You can make a concordance on more than one search term at the same time. This saves time because you don't have to make multiple concordances and it allows you to compare two or more terms quickly.

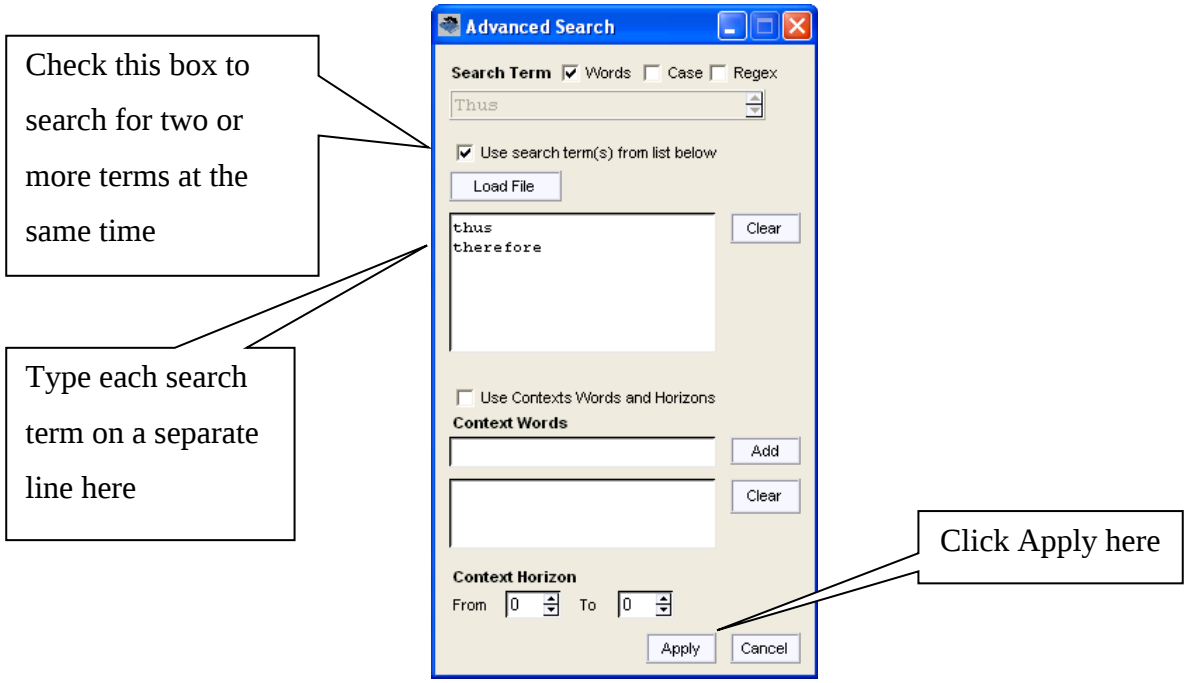
1. Click the **Advanced** box next to the Search Terms box. This takes you to the **Advanced Search** window.
2. Check the option **Use search terms from list below**.
3. Type your terms in the box below, using **one line for each term**.
4. Click **Apply** at the bottom right of the window. The **Advanced Search** window disappears.
5. At the main window, click **Start** to start the concordance in the normal way.
6. To return to normal searching, go back to the **Advanced Search** window, uncheck **Use search terms from list below** and click **Apply**.

TIP
Sort the concordance on Level 1: 0 to group the lines by each search word.

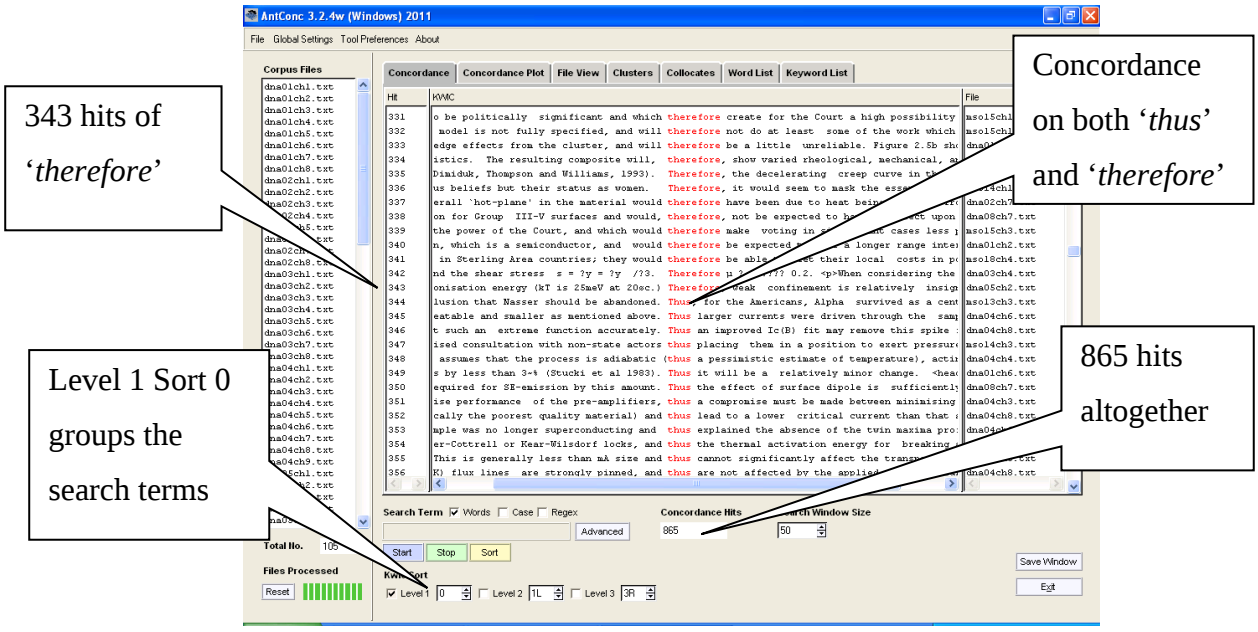
Student Question

- *Which is more frequently used in my field, 'thus' or 'therefore'?*

Screen Shot 7: Advanced Search Window for Search Terms '*thus*' and '*therefore*'



Screen Shot 8: Concordance on the Words '*thus*' and '*therefore*' at the same time



The concordance shows that in this corpus '*thus*' is more frequent than '*therefore*' because there are only 343 hits of '*therefore*' out of a total of 865 hits for both words, so '*thus*' has 522 hits.

5 Suggestions for Useful Searches

Here are some questions that other students have asked and the searches they have used to find out the answers. You can search on your expert corpus and then repeat the search on the corpus of your own writing to compare your English with what experts frequently use. This will help you to identify and deal with some of your problems.

1. What preposition should I use? Do I need a preposition?

- Search on the verb, noun or adjective that goes before the preposition. Sort on 1R to show which preposition(s) are most frequent, if any.
- Try searches on these words: *discussion*; *capable; based*.

2. What alternative verb can I use instead of 'show'?

- Make an advanced search on these terms together: *they * that; he * that; she * that; it * that* to show reporting verbs. Sort on 1R to group the verbs together.

3. How do experts present their results?

- Search on *results* and sort on Level 1:1R; Level 2: 2R and Level 3: 3R to show the verbs and patterns used.

4. Do you say 'as following' or 'as follows'?

- Make a wild card search on *follow** and sort on 1L to see which one is used.

5. How do experts discuss future research?

- Search on *further research* and *future research.* Sort on Level 1:1R; Level 2: 2R and Level 3: 3R and look at file view for more detail.

6. Can I use an adverb like 'however' at the beginning of a sentence?

- Make two separate case sensitive searches on *However* and *however*. Which one is more frequent in your corpus? Try other adverbs e.g. *thus, therefore, hence*.

7. Should I use 'a' or 'the' or neither?

- Search on the noun and sort on 1L to see which form is often used. Look at the context of the way the noun is used. Try these nouns: *research, literature, study*.

8. Which is preferred in my field: 'for instance' or 'for example'?

- Make an advanced search on the two terms *for instance* and *for example* and sort on 1R to see which one is more frequent in your field.

Good luck with your corpus searching!

One student said, "*It's like having a lovely friend with you who can advise you any time you want.*"

6 Further Information

6.1 AntConc

The concordancer is the tool most often used by students, but the *AntConc* software provides several other tools accessed by tabs at the top of the window: Concordance Plot, Clusters, Collocates, Word List and Keyword List.

You can find information on all the tools on Laurence Antony's website here:
http://www.antlab.sci.waseda.ac.jp/antconc_index.html
http://www.antlab.sci.waseda.ac.jp/software/README_AntConc3.2.4.pdf

On-line help
http://www.antlab.sci.waseda.ac.jp/software/AntConc_Help/AntConc_Help.htm

YouTube Videos covering the material in these pages:
AntConc (Ver. 3.2) - Getting Started
AntConc (Ver. 3.2) - Tutorial 1: Concordance Tool - Basic Features
AntConc (Ver. 3.2) - Tutorial 2: Concordance Tool - Advanced Features
AntConc (Ver. 3.2) - Tutorial 4: File View Tool - Basic Features

Laurence Antony is gratefully acknowledged for his permission to use the software and screen shots in this material.

6.2 Courses Using Corpora at the Language Centre

There are two types of course using corpora and *AntConc* run by the Language Centre. Both are for non-native speakers of English only.

1. For DPhil Students only: *Edit your Thesis with Corpora*

This course is designed to help students who are writing up their thesis. It is a 6-week (12 hour) course that runs every term starting in Week 3.
The syllabus is available here:

http://www.lang.ox.ac.uk/files/Syllabi/Doctoral%20Training%20Syllabus%202013.doc

2. For all Students: *Writing in your Field with Corpora*

This course is designed to help students examine the ways in which writers in their field use language to perform academic functions such as making arguments and citing other researchers. It is part of the Academic Writing Programme and runs in Trinity Term only.

The syllabus is available on the Language Centre website here:

http://www.lang.ox.ac.uk/courses/english.html

6.3 Introduction to Corpus Linguistics Course

The IT Learning Programme offers an introductory course in corpus linguistics, which consists of 6 lunch-time sessions. The details are here:

https://weblearn.ox.ac.uk/portal/hierarchy/central/oucs/itlp_courses/corpus_ling