

## **The Prime Machine HD Guide** ***I want to ... revise my thesis***

If you are a final year undergraduate working on your dissertation or a postgraduate student writing up your thesis, you will probably have been working with the text of your project over many weeks or months. Good academic writing involves reviewing, editing and revising your thesis in order to correct typos and language errors; to improve clarity and fluency; and to make the progression of your ideas sharp and well-organized. While there are tools to help you notice issues with grammar and nothing can replace the time you should spend reading through from start to finish, working with a corpus of your own writing and the texts of your sources can help you notice expressions to be retuned or reworded which you may otherwise overlook. This guide will explain the steps you can take to build a DIY corpus of your own thesis chapters in The Prime Machine HD corpus tool and then explore this corpus with reference to a corpus of your sources or a readymade corpus of academic English from the server. You won't edit your thesis using tPM. But you can search through your text in different ways to help you think about what you could revise.

Steps to complete:

1. Before you start – **make a backup of your work!** Yes. Now!
2. Import the chapters of your thesis (ideally chapter by chapter or as a complete text).
3. Import the articles you read and cited in your thesis.
4. Use the corpus tools to explore different aspects of your own writing.

This approach can help you:

1. Explore your own writing in a less linear way; seeing the links you've made between paragraphs and chapters through repetition of words and phrases.
2. Find examples of signposting language and the words you've used to structure your text, and then compare these strategies with those of published academic writers.
3. Check the language you have used around key terminology to see whether your use of verbs and noun phrases match your sources.
4. Explore your reporting verbs, to ensure you have successfully communicated your attitude and stance as you introduce the ideas and research of others.

What you'll need to get started:

- The Prime Machine HD for Windows, macOS, iPad, iPhone or Android (available free from <https://www.theprimemachine.net/> )
- Your dissertation as plain text, RTF, Word or PDF.
- Your sources as plain text, RTF, Word or PDF.
- Patience, enthusiasm and an open mind!

### **Looking at your dissertation through linguistic spectacles**

You don't need to be a student of linguistics to be able to analyse the patterns of use of language (but of course if you've studied linguistics you can draw on your linguistic knowledge). When you submit your dissertation, you are essentially showing to your professors, the institution and the world that you are becoming a member of an academic community – a community which communicates using academic conventions and through academic language. The way you structure your thesis demonstrates knowledge of the structure of academic writing. The way you introduce ideas in your literature review not only demonstrates your knowledge of the research of others, but also reveals your attitudes and confidence in their work. Your use of specialist terminology reveals your experience reading and hearing these words and phrases as they are applied in your specialist

discipline. Just as we can say you are what you eat, we can say the wording of your thesis reveals your specialist academic knowledge – your knowledge is what you write.

## Are you ready to become a language detective?

To make the most of the approach described here, you will need to try to think about your thesis and the texts that you have read as a detective might look at clues. As well as being an amazing expert in detection, Sherlock Holmes is also remembered in the stories as being an expert in disguise.

- As a detective, can you look at a hundred examples of a word or phrase and pick out some of the patterns which occur?
- As a detective, can you look at the context and co-text of examples from a corpus to see deeper meanings and typical uses of words?
- As an expert in disguise, can you dress-up your own writing, so as to blend into the norms and expectations of expert language users?

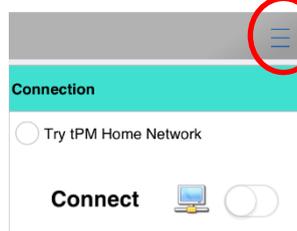
## Getting started

The best place to get The Prime Machine HD (tPM) is from an official store. It is free!



Windows and Android users can also download the App directly from the website:  
<https://www.theprimemachine.net>

When you first use tPM, you almost certainly will want to connect to the server to access pre-prepared corpora and resources. There are two main views for the search screen – Simple Mode and Full Mode. The Full Mode includes additional tabs and features for corpus research and DIY corpus work. In this guide, you will need to use Full Mode.



The main 'hamburger' or 蒸笼 menu in the top-right corner allows you to connect and change mode.

## Simple Mode



## Full Mode



## Step 1: Building a DIY corpus of the chapters of your thesis

Before you start importing the chapters of your thesis, you should think about whether splitting the long file into chapters would be helpful, and then think about the filenames and the order of the files.

If you load your whole thesis as a single text, you will still be able to complete most of the analyses in Step 3, but you won't be able to explore the chapters separately. If you do have separate files for each chapter, make sure the filenames are neat and consistent. For example, rename the files "Chapter 1 Introduction", "Chapter 2 Literature Review", etc.

If you are using iPad, iPhone or Android, use your device's file manager to make a zip file of all the chapters first. On desktop platforms, you can load multiple files in one operation.

Screenshots of the procedure on all platforms are on pages 4-7. Remember to SAVE THE DIY CORPUS.

## Step 2: Building a second DIY corpus of your sources

The process for building a second DIY corpus containing as many of your sources as possible is precisely the same. Before you import the files, change the filenames so they will show meaningful information when you look at the results later. For example, if you rename the file using the author's name and year of publication, you will be able to recognise each hit much more easily when you do searches.

You can also group texts into categories for your DIY corpus. These will be displayed when you look at the concordance cards (see page 11). Categories can be loaded automatically if you put texts into subfolders and give the subfolder a meaningful name. For example, you could use subfolder names such as "Theoretical framework" or "Early studies". These will also be displayed when you look at the concordance cards, but otherwise you can just use a single name for your whole corpus.

Remember to save time by loading all the sources in one operation. Adding one text to an existing corpus will be slower than loading all the texts together, as tPM will need to re-index all the words and combinations of words. On Windows/macOS you can load multiple files by selecting them all together and dragging them into the drop-zone, selecting multiple files from the open file dialog, using the button to open all the files in a single folder, or by zipping the files first. On iPad, iPhone and Android, the only way to open multiple files in one operation is to zip them first and then use "share" or "open with" to import the zip into tPM.

When you have built your DIY corpus of your sources, save it on your device.

## File formats

You can import text from a variety of file formats: PDF, DOCX, DOC, RTF, TXT, PPTX, PPT and EPUB.

However, some PDF files may not be compatible because they contain scanned images of text or have other restrictions.

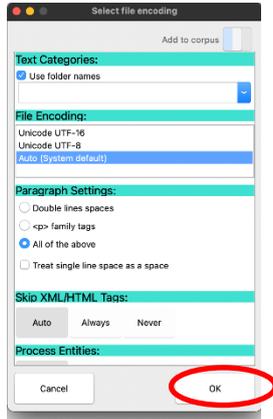
Your aim should be to load a good selection of texts from your sources; you don't need to include everything.

If there is a source you really want to include, but it doesn't load properly in tPM, you can consider using the free desktop app tPMCrafty (also available from The Prime Machine website) or use other tools to convert your PDF to plain text first.

tPMCrafty can help add spaces to the ends of lines, alter the spacing between paragraphs and split a text into smaller files.

## MacOS

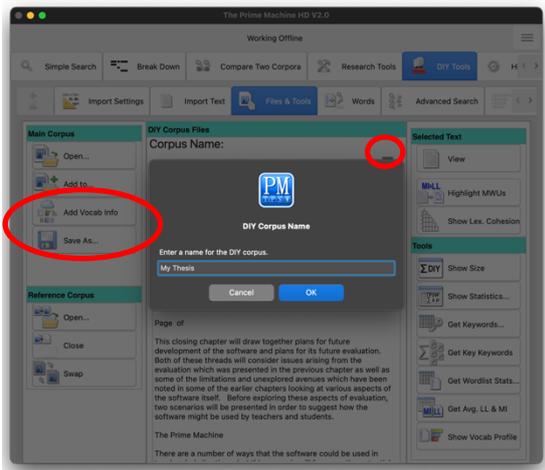
When the App starts, make sure Full Mode is selected and go to the DIY Tools – Import Text tab.



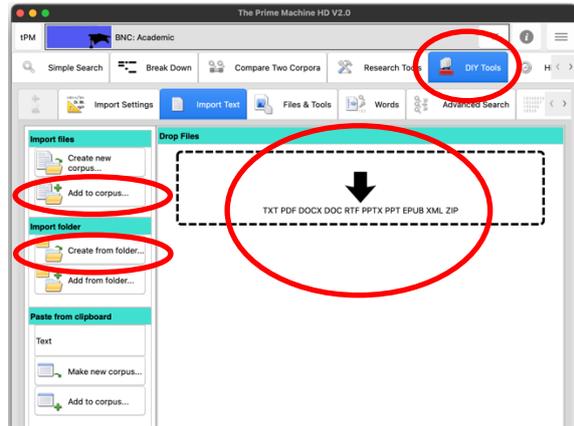
Most default settings will work well. Alter the file encoding or paragraph settings if working with plain text files and you get unexpected results.



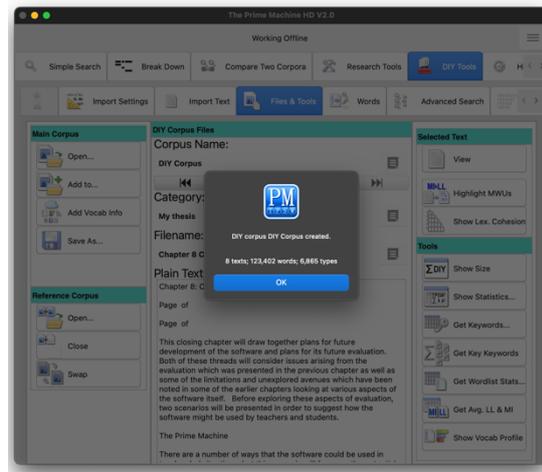
Wait patiently while the text is extracted, the sentences and paragraphs are organised and the words and combinations of words are indexed.



From the Files & Tools Tab you can rename the corpus and then save it using the "Save as..." button.



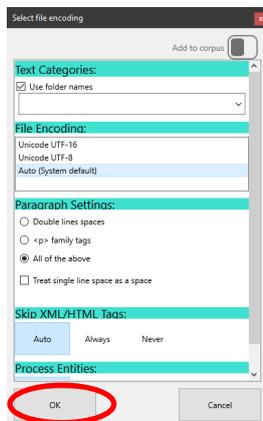
Drag and drop the files into the drop zone; or use the "Create new corpus..." button to select one or more files inside a folder; or choose "Create from folder..." to import an entire folder of files.



When it is finished, you will see the size in texts, words and types.

## Windows

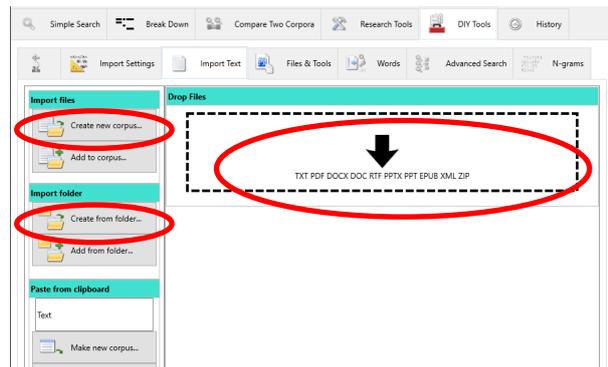
When the App starts, make sure Full Mode is selected and go to the DIY Tools – Import Text tab.



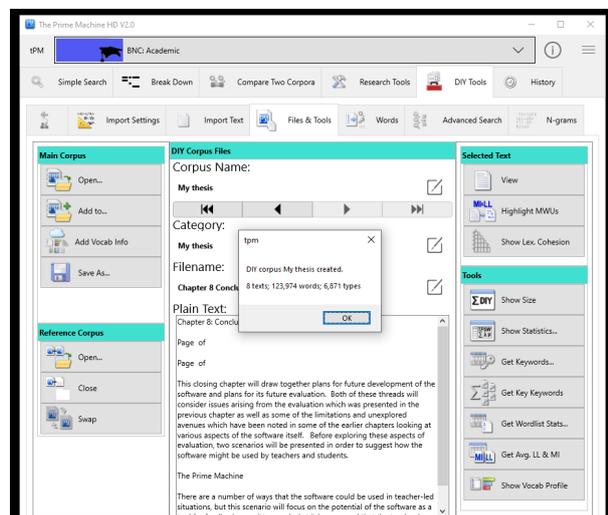
Most default settings will work well. Alter the file encoding or paragraph settings if working with plain text files and you get unexpected results.



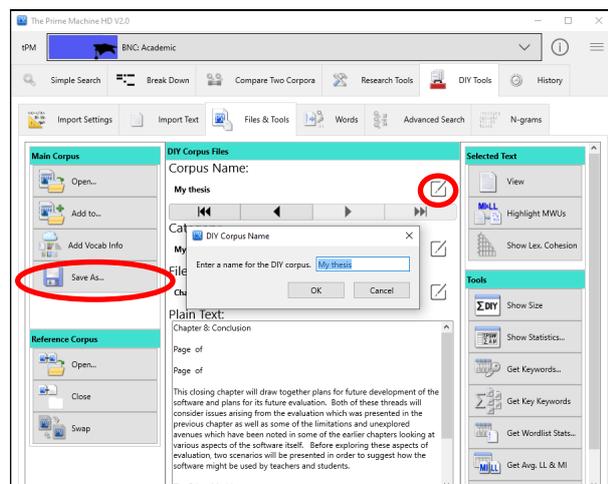
Wait patiently while the text is extracted, the sentences and paragraphs are organised and the words and combinations of words are indexed.



Drag and drop the files into the drop zone; or use the "Create new corpus..." button to select one or more files inside a folder; or choose "Create from folder..." to import an entire folder of files.

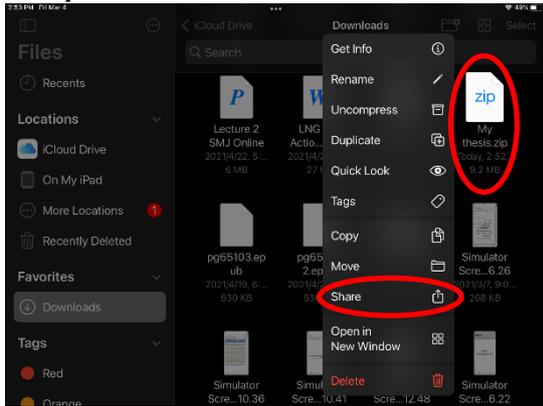


When it is finished, you will see the size in texts, words and types.

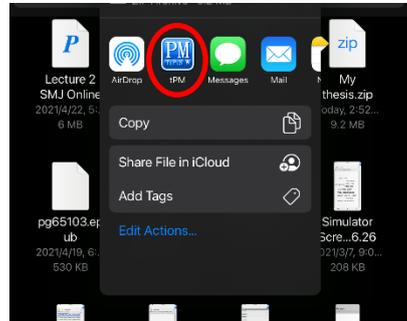


From the Files & Tools Tab you can rename the corpus and then save it using the "Save as..." button.

### iPad / iPhone

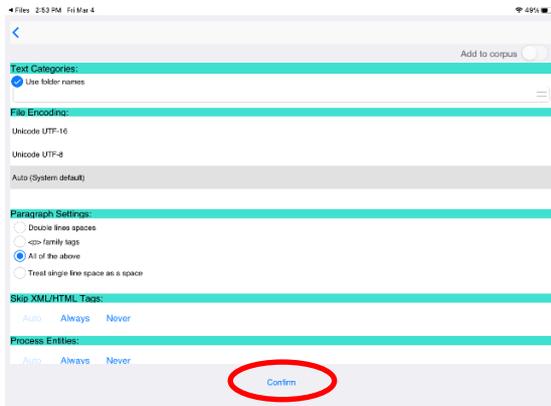


**On mobile platforms, tPM only has access to your files when you share them from another app.**

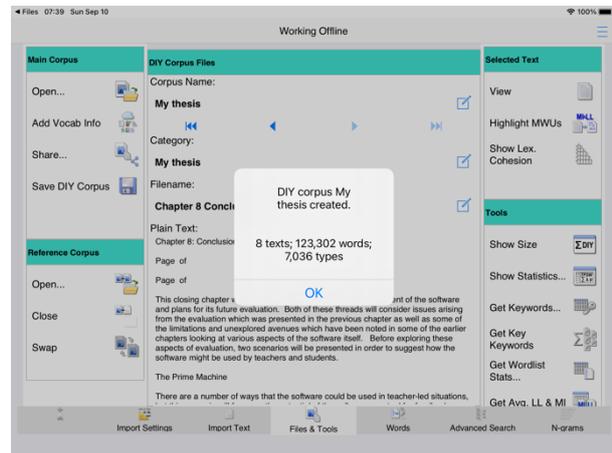


Use the "Files" App to find the zip file (or single document) you want to import into tPM. Long tap on it and then choose "Share".

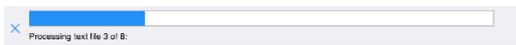
Find tPM on the list of apps and tap on it.



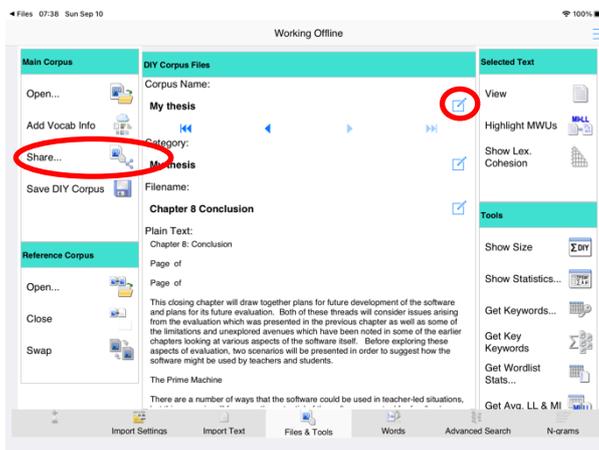
Most default settings will work well. Alter the file encoding or paragraph settings if working with plain text files and you get unexpected results.



When it is finished, you will see the size in words and types.

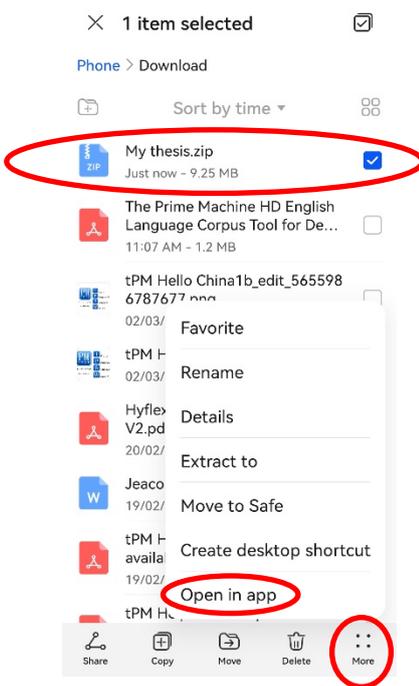


Wait patiently while the text is extracted, the sentences and paragraphs are organised and the words and combinations of words are indexed.



From the Files & Tools Tab you can rename the corpus and then save it using the "Save" button.

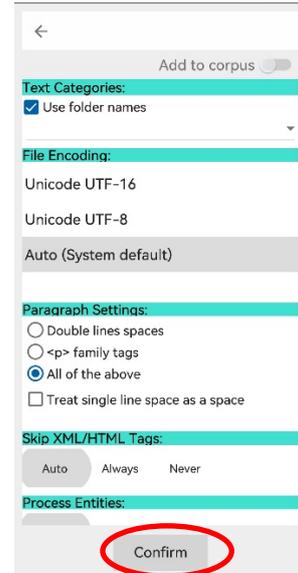
### Android



**On mobile platforms, tPM only has access to your files when you share them from another app.**

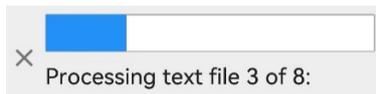


Find tPM on the list of apps and tap on it.

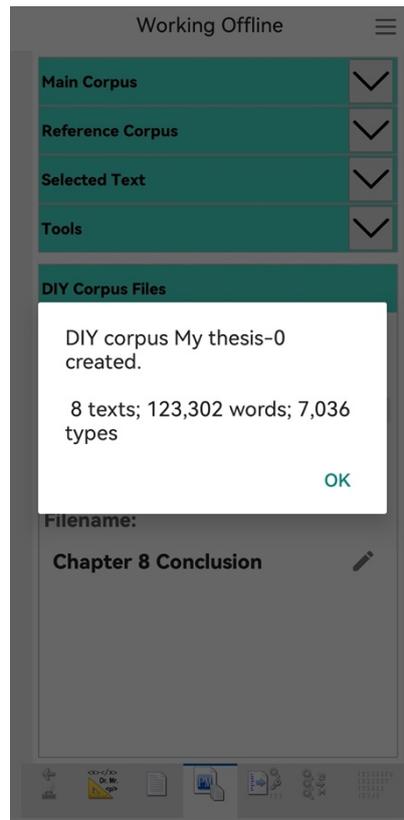


Most default settings will work well. Alter the file encoding or paragraph settings if working with plain text files and you get unexpected results.

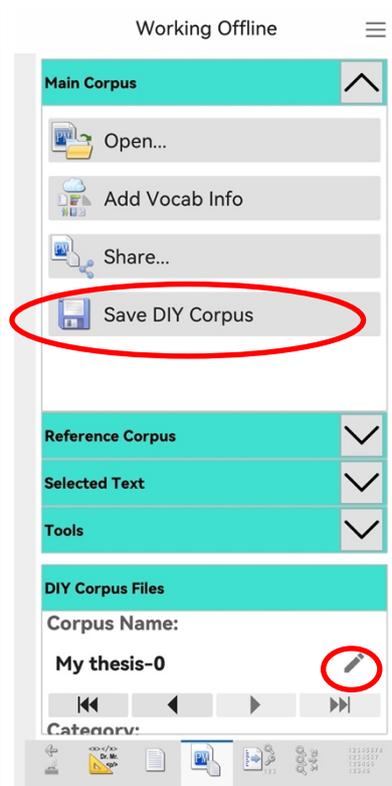
Use the "Files" App to find the zip file (or single document) you want to import into tPM. Long tap on it and then choose "More" and "Open in app"



Wait patiently while the text is extracted, the sentences and paragraphs are organised and the words and combinations of words are indexed.



When it is finished, you will see the size in texts, words and types.



From the Files & Tools Tab you can rename the corpus and then save it using the "Save" button.

### **Step 3: Analysing your writing**

Now you have your thesis and your sources loaded as DIY corpora, you can explore your texts using data driven methods as well as specific searches. The data driven methods (3.1 – 3.3) use repetitions or combinations of words to compare your thesis with a reference corpus (your sources or a readymade corpus like the BNC: Academic). These data driven methods may help you notice things you may not notice otherwise. Specific searches (3.4 - 3.6) use a list of words or a built-in procedure to help you explore aspects we can predict could be important. You could get the same information using other tools or manually counting, but the corpus tool makes it smooth and easy.

#### **3.1 Finding the core terminology (keywords and key keywords)**

##### **What does it do?**

The keyword method goes through your corpus, counting the frequencies of each different word. It then compares these frequencies with a reference corpus. The words which are important in a text (or collection of texts) are likely to be repeated many times. But some words are generally more common than others – consider the number of times we use the word *of* every day, compared to the number of times we might use someone's name or a less common word like *university*. In texts about a university course or research on a university, we would expect the word *university* to occur many more times than it would do normally. So basically, the keyword procedure uses the list of words in your thesis and compares each word's frequency to the frequency in the reference corpus. Then it performs a statistical test to help move the most surprising or notable differences in frequency to the top.

The key keywords procedure is the same kind of process, but it operates text by text and then gives an overall summary of words which occur in at least two keyword lists in your corpus. If you have loaded your thesis as separate chapters, you can use this function to see which words are prominent in two or more chapters. If you only have one file in your DIY corpus, you cannot use this function.

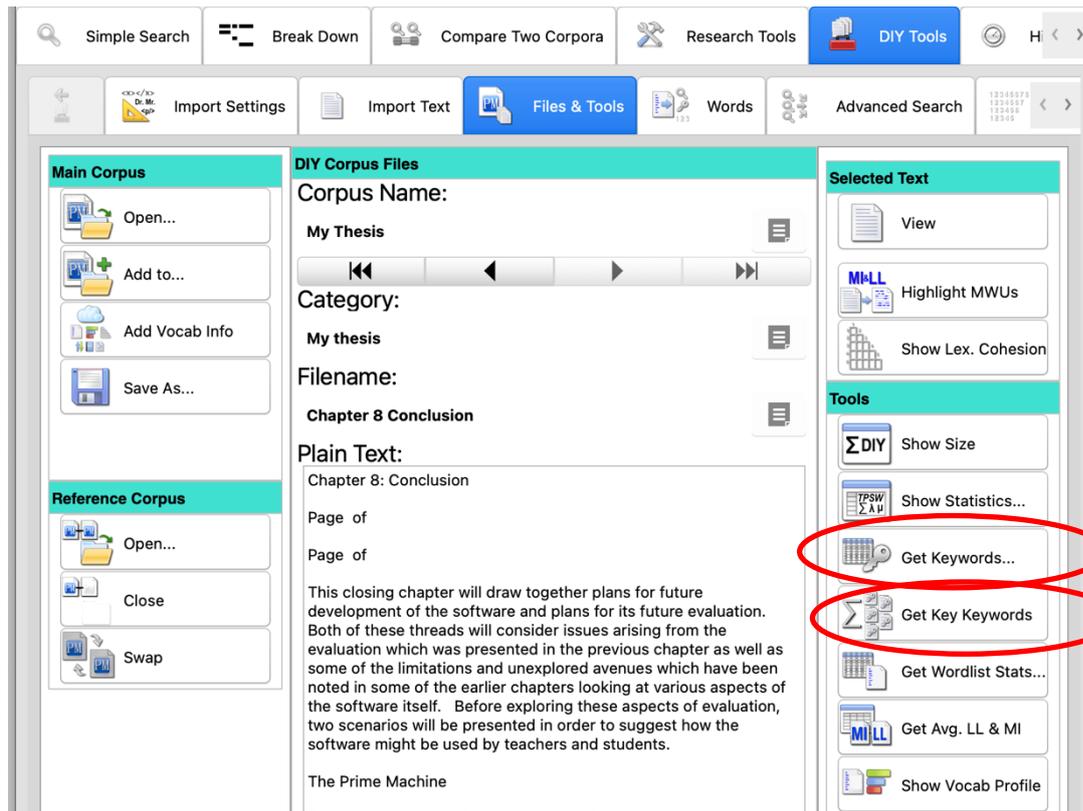
Essentially, the reference corpus becomes a baseline for the expected frequencies of English words. Words used more often than would be expected (based on this baseline) become keywords for your thesis. These keywords are a different kind of key words researchers choose for themselves when they list them under an abstract (but there may be some overlaps).

##### **How do I do it?**

The first thing to decide is whether you want to generate keywords comparing your thesis with your sources or with a general academic corpus. There are benefits to doing both, so you can repeat the process and you will get slightly different results. If you want to have the option to get keywords from a readymade corpus, you need to connect to the tPM server first. Do this using the 'hamburger' or 蒸笼 menu (see page 2).

If you want to use your DIY corpus of your sources as the reference corpus, you need to have loaded your thesis as the Main DIY corpus and the sources as the DIY reference corpus. You can save and load DIY corpora from the Files & Tools tab.

If you are connected to the server and have a DIY reference corpus loaded, you'll be asked which one you want to use.



To get keywords or key keywords, all you need to do is go to the File and Tools menu (under DIY Tools) and click or tap the button. If you are using a readymade corpus as a reference corpus, the app will send the list of words and their frequencies from your thesis and compare them on the server. None of your data is stored on the server.

If you are using a second DIY corpus as a reference corpus, the app does not need to access the server for these functions.

The screenshot shows a table titled 'Keywords for My thesis compared against BNC: Academic'. The table has columns for Word, Study\_Freq, Study Per, and LL Bayes. A context menu is open over the table, showing options like Copy text, Copy image, Save image..., Save sheet..., Save as PDF..., Show/Hide columns, and View Search History.

Word	Study_Freq	Study Per	LL Bayes
1 corpus	527		17 Very strong evidence
2 word	635		83 Very strong evidence
3 concordance	296		20 Very strong evidence
4 corpora	252		81 Very strong evidence
5 collocations	244		76 Very strong evidence
6 collocation	231		13 Very strong evidence
7 words	586		96 Very strong evidence
8 software	360		27 Very strong evidence
9 learners	249		88 Very strong evidence
10 students	399		12 Very strong evidence
11 text	364		99 Very strong evidence
12 tab	148		71 Very strong evidence
13 bnc	120		25 Very strong evidence
14 tags	129		34 Very strong evidence
15 png	109	0.94	0 1102.03 Very strong evidence
16 lines	251	2.16	1,977 0.11 ≥ 10 994.33 Very strong evidence
17 priming	117	1.01	68 0.00 ≥ 100 940.46 Very strong evidence
18 server	94	0.81	4 0.00 ≥ 100 917.00 Very strong evidence
19 texts	186	1.60	987 0.05 ≥ 10 867.28 Very strong evidence
20 database	165	1.42	671 0.04 ≥ 10 846.26 Very strong evidence
21 sentence	226	1.95	2,488 0.14 ≥ 10 760.63 Very strong evidence
22 user	170	1.46	1,020 0.06 ≥ 10 755.74 Very strong evidence
23 xml	72	0.62	0 0.00 ≥ 1 727.95 Very strong evidence
24 claws	80	0.69	17 0.00 ≥ 100 719.01 Very strong evidence
25 results	282	2.43	5,087 0.28 ≥ 5x 705.46 Very strong evidence

When the results are ready, you will see a table of words and other statistics.

Just like with all tables of data in tPM, you can copy, save or share the results as an image, a spreadsheet file or a PDF document.

Simply double-click, right-click or long tap to bring up the menu.

On Windows and macOS, you can copy or save.

On mobile platforms, you can share the pictures, spreadsheets or PDFs to other apps.

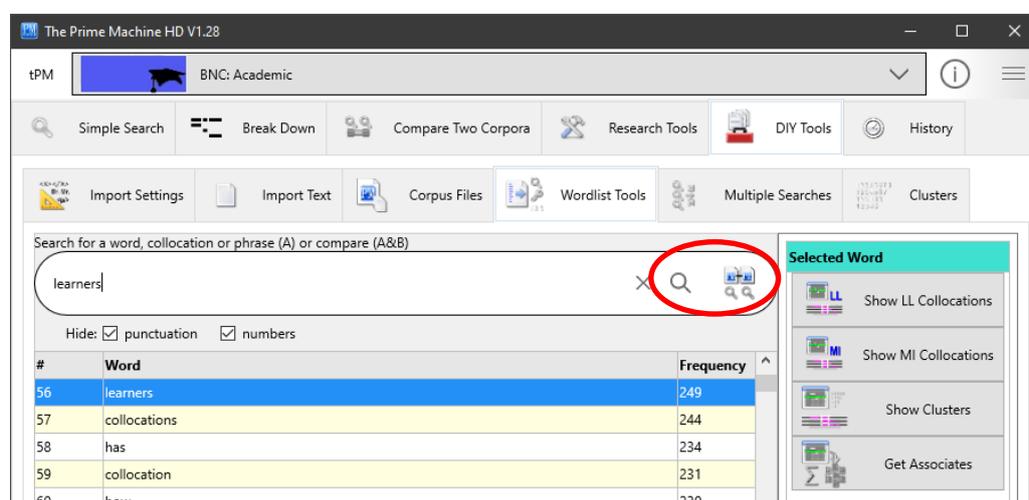
## What does it mean?

Although we may be told by our English teachers not to repeat words in our essays, for an extended text like a thesis, it is expected that your core terminology will be repeated again and again. Words which don't occur in the reference corpus will have a sun symbol. Others will show an arrow and a number to indicate the magnitude of difference. In most cases, you will just want to note down some of the keywords and use them to explore the patterns of usage in your own writing, compared to the sources or a readymade corpus.

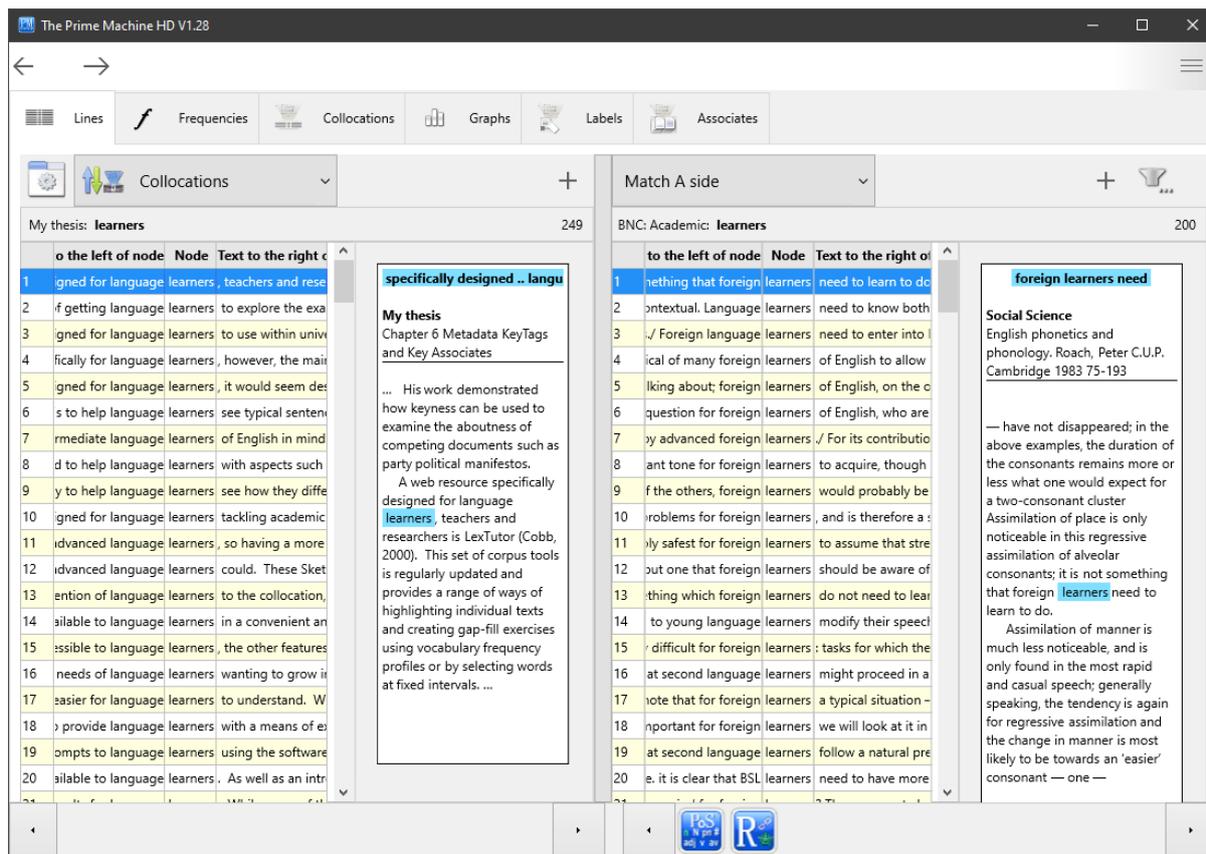
1. Look out for spelling mistakes. If you've misspelled a word more than once it will probably show up here!
2. Take a few words of core terminology from your list and use them for concordance line searches to explore the patterns of use (see below).
3. Look out for signposting words in the keyword list – and look at them even more seriously if they are key keywords! Normally, we don't repeat signposting words very often in a single text, and because you've used another corpus as a baseline, only if you have used words like *however* or *nevertheless* or *moreover* many many times would they appear as keywords.
4. Don't be surprised to see some of the names of the main influences on your research appear as keywords. But it might be interesting to see if you've repeated someone's name many times when really their research is more tangential. This is probably something to discuss with a supervisor, rather than simply edit out for language reasons.

## What should I do with keywords and key keywords?

Copy some of the core terminology and paste (or type) into the search box on the Wordlist Tools tab. Then click or tap the search button to only look inside your thesis or the compare button the display hits in your thesis on the left with a reference corpus on the right.



From this screen you can look up individual words and strings of words (with no other words between). You can do more complicated searches on the Advanced Search tab (see pages 24 and 28).

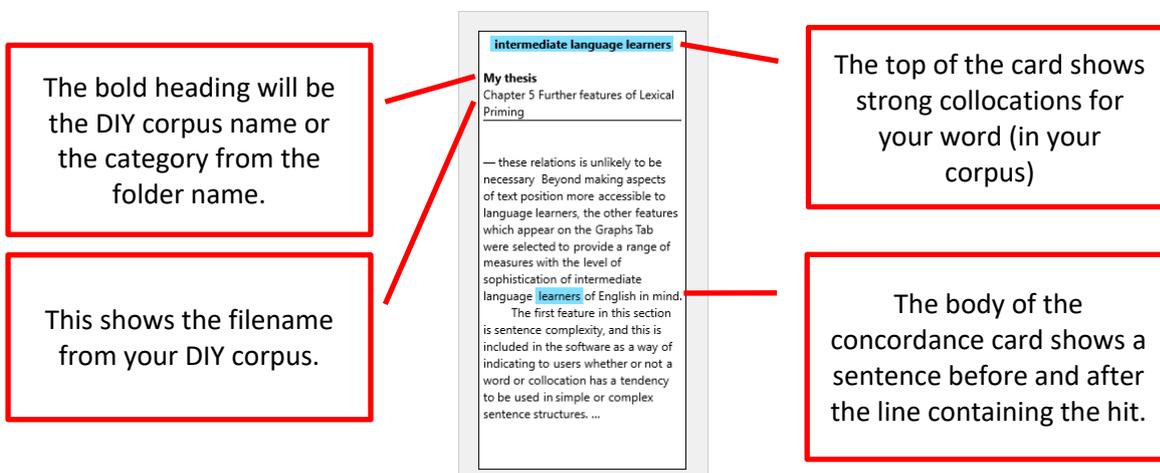


There are many things you can do when you look at concordance lines.

The default sorting method in tPM is using Collocations. This means the hits will appear in a different order from the order in your thesis. You'll see the lines with the strongest collocations at the top.

You can change the sorting method to explore the results in different ways. For example, Text Order will present your results in the order as they appear in your thesis (just like the order if you use the "Search" function in your word processor).

As you select a line, you will see the concordance card with the wider context in a white box to the right.



You can look to see if the combinations of words you have used are similar or different from your sources. You can also see how these words are used in a more general corpus.

Concordance lines and cards can be copied, saved or shared. Double-click, right click or long tap on the lines or the cards to open the menu.

Take some of the words from your core terminology and ask yourself:

- Are the verbs I have used with nouns similar to those in the reference corpus?
- Have I used noun phrases for the core terminology using the same combinations of nouns, adjectives and other elements?
- Can I spot any grammatical patterns which seem to be different for the word in my corpus, compared to the reference corpus?

### 3.2 Finding repeated chunks (clusters or n-grams)

#### What does it do?

Another way of looking at the text of your thesis is to use the corpus tool to find repeated strings of words. In corpus linguistics and computer science these are known as clusters or n-grams. In tPM, the N-grams tab under DIY Tools allows you to extract lists of strings of words of between 3 and 8 words in length. These n-grams will be continuous chunks in the text with no words in-between. Sentence by sentence, strings of words will be generated, starting with the first word and moving along 8 words, then starting with the second word and moving along 8 words, and so on. My own thesis contains this sentence in Chapter 5:

- This shows the proportion of concordance lines where the word is in the Theme or Rheme of the sentence.

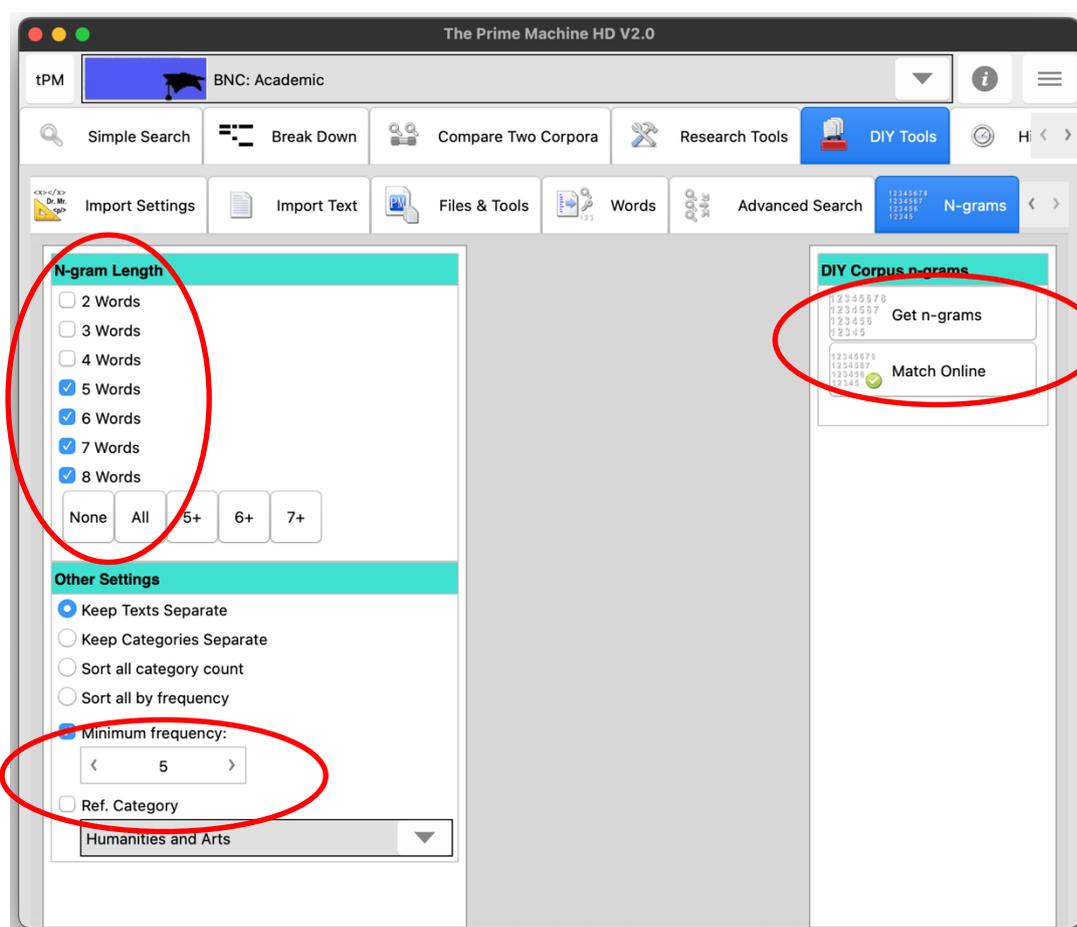
The computer will generate 12 strings of 8 words from this single sentence as follows:

This →	This shows the proportion of concordance lines where the word is in the Theme or Rheme of the sentence.
shows →	This shows the proportion of concordance lines where the word is in the Theme or Rheme of the sentence.
the →	This shows the proportion of concordance lines where the word is in the Theme or Rheme of the sentence.
proportion →	This shows the proportion of concordance lines where the word is in the Theme or Rheme of the sentence.
of →	This shows the proportion of concordance lines where the word is in the Theme or Rheme of the sentence.
concordance →	This shows the proportion of concordance lines where the word is in the Theme or Rheme of the sentence.
lines →	This shows the proportion of concordance lines where the word is in the Theme or Rheme of the sentence.
where →	This shows the proportion of concordance lines where the word is in the Theme or Rheme of the sentence.
the →	This shows the proportion of concordance lines where the word is in the Theme or Rheme of the sentence.
word →	This shows the proportion of concordance lines where the word is in the Theme or Rheme of the sentence.
is →	This shows the proportion of concordance lines where the word is in the Theme or Rheme of the sentence.
in →	This shows the proportion of concordance lines where the word is in the Theme or Rheme of the sentence.

The shorter n-grams will overlap with the longer ones. For example, the 8-gram (8 word cluster) "shows the proportion of concordance lines where the" will contain 5-grams such as "shows the proportion of concordance", and "the proportion of concordance lines".

### How do I do it?

When looking at other kinds of text, it is often useful to only count n-grams with a minimum frequency of 5. But in a thesis it is probably a good idea to change this minimum frequency to 2 – that means if you've used the same string of words anywhere in your thesis it will show up on the list. If you are just looking for very long strings of repeated words, you can just choose 7-grams and/or 8-grams. To explore some other common ways you introduce ideas, choose shorter n-gram (between 3 and 5 words), and consider increasing the minimum frequency (perhaps between 3 and 5).



If you want to explore strings of words used in academic language more generally, you can also use the Match Online Corpus button. This will send your thesis to the server cluster by cluster and try to find matches in the readymade corpus. That way, you can get a sense of which strings of words are common in general academic writing and which strings of words are more specialized.

**Note: Using "Match Online Corpus" means chunks of your thesis will be transmitted over the internet. Efforts are made to make this secure, but you should be mindful of the risks of transmitting these kinds of data.**

### What does it mean?

Many of the n-grams will be noun phrases of important concepts or perhaps formulaic language you need in captions for figures and tables. Having repeated strings of words is not a problem in itself (unless like a couple of my own students you have moved a section of your thesis and forgotten to delete it from the original location!).

If you have loaded your thesis with each chapter in a separate file, the "texts" column tells you the number of chapters in which each cluster appears.

Looking at line 3 in the table below, we can see I seem to have a favourite expression for introducing chapters. The cluster "the purpose of this chapter is to introduce" occurs 3 times in a total of 3 chapters; three different chapters have this same formulaic introduction. This isn't a problem in itself (nobody noticed this in my own thesis!). But these signposting phrases are another aspect you can focus on when exploring the n-grams from your thesis.

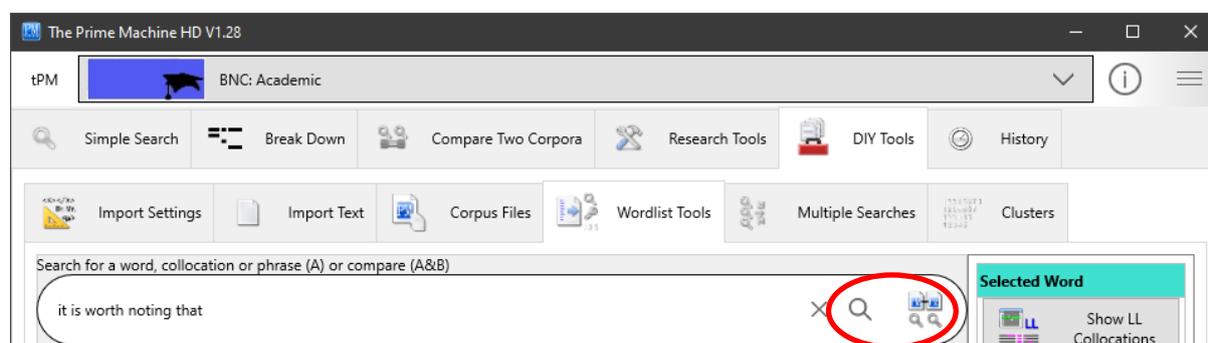
Cluster	Category	Length	Freq.	Per Thousand	Texts
1 shows the proportion of concordance lines where the	My thesis	8	5	0.04	1
2 shows the proportion of concordance lines which are	My thesis	8	5	0.04	1
3 the purpose of this chapter is to introduce	My thesis	8	3	0.03	3
4 tendencies of words and collocations to occur in	My thesis	8	3	0.03	2
5 a good idea to look at the concordance	My thesis	8	3	0.03	1
6 good idea to look at the concordance lines	My thesis	8	3	0.03	1
7 idea to look at the concordance lines to	My thesis	8	3	0.03	1
8 more often than would be expected by chance	My thesis	8	3	0.03	1
9 more than half of the occurrences of the	My thesis	8	3	0.03	1
10 this shows the proportion of concordance lines which	My thesis	8	3	0.03	1
11 to look at the concordance lines to see	My thesis	8	3	0.03	1
12 yet more than half of the occurrences of	My thesis	8	3	0.03	1
13 about language based on hoey 's theory of	My thesis	8	2	0.02	2
14 according to some of the features of lexical	My thesis	8	2	0.02	2
15 based on hoey 's theory of lexical priming	My thesis	8	2	0.02	2
16 card for the currently selected row on the	My thesis	8	2	0.02	2
17 concordance lines in which the node word is	My thesis	8	2	0.02	2
18 for a message to be sent from the	My thesis	8	2	0.02	2
19 for the currently selected row on the lines	My thesis	8	2	0.02	2
20 goes some way to providing a platform for	My thesis	8	2	0.02	2
21 hoped that the prime machine goes some way	My thesis	8	2	0.02	2
22 in contexts related to features of lexical priming	My thesis	8	2	0.02	2
23 is hoped that the prime machine goes some	My thesis	8	2	0.02	2
24 it is hoped that the prime machine goes	My thesis	8	2	0.02	2
25 language based on hoey 's theory of lexical	My thesis	8	2	0.02	2

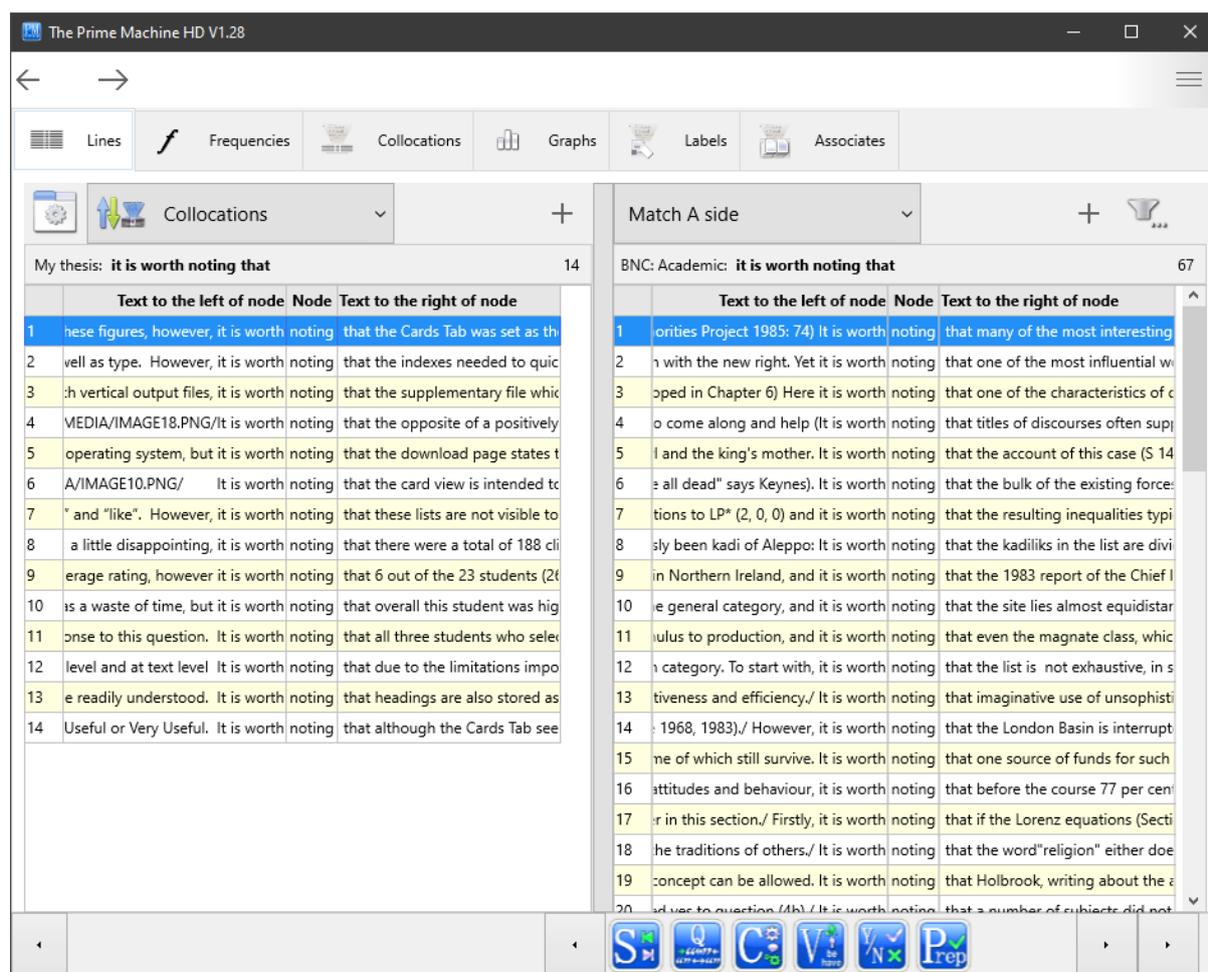
Like all the other data tables in tPM, you can export the results as a spreadsheet if you wish. Then you could classify the n-grams according to their role or function. For example, in the table of 5-grams (5 word clusters) below, I could categorise formulaic language for introducing examples (1, 4, 5, 6, 9, 10, 11, 15, 16, 17 and 19), drawing attention (3, 13, 18, 22 and 24), and some important influences or topics (2, 12, 14 and 23).

	Cluster	Category	Length	Freq.	Per Thousand	Texts
1	of the occurrences of the	My thesis	5	17	0.15	1
2	the theory of lexical priming	My thesis	5	14	0.12	6
3	it is worth noting that	My thesis	5	14	0.12	5
4	shows the proportion of concordance	My thesis	5	13	0.11	1
5	the proportion of concordance lines	My thesis	5	13	0.11	1
6	the occurrences of the word	My thesis	5	11	0.09	1
7	it would be possible to	My thesis	5	10	0.09	4
8	either side of the node	My thesis	5	10	0.09	3
9	the overall proportions of tokens	My thesis	5	10	0.09	1
10	can be seen in figure	My thesis	5	9	0.08	4
11	as can be seen in	My thesis	5	9	0.08	3
12	used in the prime machine	My thesis	5	8	0.07	2
13	it should be noted that	My thesis	5	7	0.06	4
14	in the prime machine is	My thesis	5	7	0.06	3
15	can be seen in table	My thesis	5	7	0.06	1
16	proportion of concordance lines where	My thesis	5	7	0.06	1
17	shows the overall proportions of	My thesis	5	7	0.06	1
18	is worth noting that the	My thesis	5	6	0.05	5
19	it can be seen that	My thesis	5	6	0.05	5
20	the way in which the	My thesis	5	6	0.05	5
21	at the end of the	My thesis	5	6	0.05	4
22	it is clear that the	My thesis	5	6	0.05	4
23	the design of the software	My thesis	5	6	0.05	4
24	it could be argued that	My thesis	5	6	0.05	2
25	it would have been possible	My thesis	5	6	0.05	2

### What should I do with n-grams?

You can find out more about how you have used n-grams and also explore these n-grams in a reference corpus. For DIY corpora, on the Wordlist Tools tab, you can only search for n-grams of up to 5 words.





You could look at concordance lines from your own thesis and compare them with concordance lines from a readymade corpus or your DIY corpus of sources.

In the image above, we can see in my own thesis lines 4 and 5 have the filenames of the images (IMAGE 18.PNG & IMAGE 10.PNG). This shows that in two cases, I have begun a paragraph below a figure with the words "It is worth noting". We can also see that I have a tendency to use "it is worth noting" to introduce constraints or limitations (lines 2, 7 and 12).

### 3.3 Finding word combinations which blend in or stand out (collocation highlighting)

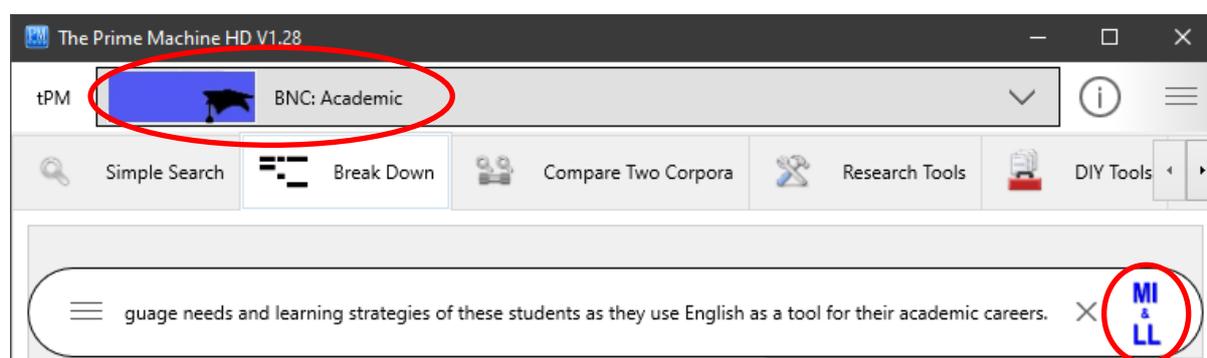
#### What does it do?

Collocations are the conventional way of expressing ideas – the relationships between words which make combinations sound natural. They are sometimes called "word partnerships" in dictionaries and English language textbooks. When members of a community want to express an idea, they tend to use the combinations of words that have been used before; where there is a choice of verbs with similar meanings, only one verb seems the natural choice; where there could be several adjectives with a similar meaning, only one seems to fit a specific situation. Collocation strength can be calculated by comparing the frequency of a specific combination against the frequencies of each component separately. A computer can give a sentence a collocation score by checking each combination of words against a reference corpus. Essentially, when you use this function, tPM will send pairs of words which are one, two, three and four words apart and check these in the readymade corpus.

Several different collocation scores are then combined to generate a colour-code for the words in the sentence. Words which form strong collocations with their neighbours will tend towards being shaded green, while words which stand apart from their neighbours will tend towards grey.

### How do I do it?

There are two ways to use the collocation highlighting tool in tPM: using the search box on the Break Down tab or by sending an entire text from the DIY corpus using a button on the Files & Tools tab under DIY Tools. With either option, the first thing to consider is what you want to use as the reference corpus. If your thesis is in an academic discipline which matches one of the specialist Hindawi corpora, you may wish to select that. Otherwise, the BNC: Academic can be a good source for academic collocations. And the British National Corpus (complete) could be good if you want to match general English.



On the Break Down tab, type or paste the sentence into the search box and click the MI & LL button. The combinations of words will be sent to the server and the results will appear as shown below.

Line	Score	MI	DeltaP	LL	Sentence
1	6.55	<a href="#">100</a>	<a href="#">100</a>	<a href="#">50</a>	as the number of students studying degree subjects through the medium of english grows , it is becoming increasingly important to understand the language needs and learning strategies of these students as they use english as a tool for their academic careers

As well as viewing the results and noting down some combinations to explore later, the numbers in the columns for MI, DeltaP and LL have links, so you can see all the matching collocations and their scores.

In the table below, the LL score of 50 has been clicked, revealing a table of Log-likelihood collocations found in the BNC: Academic corpus which match combinations of words in our sentence.

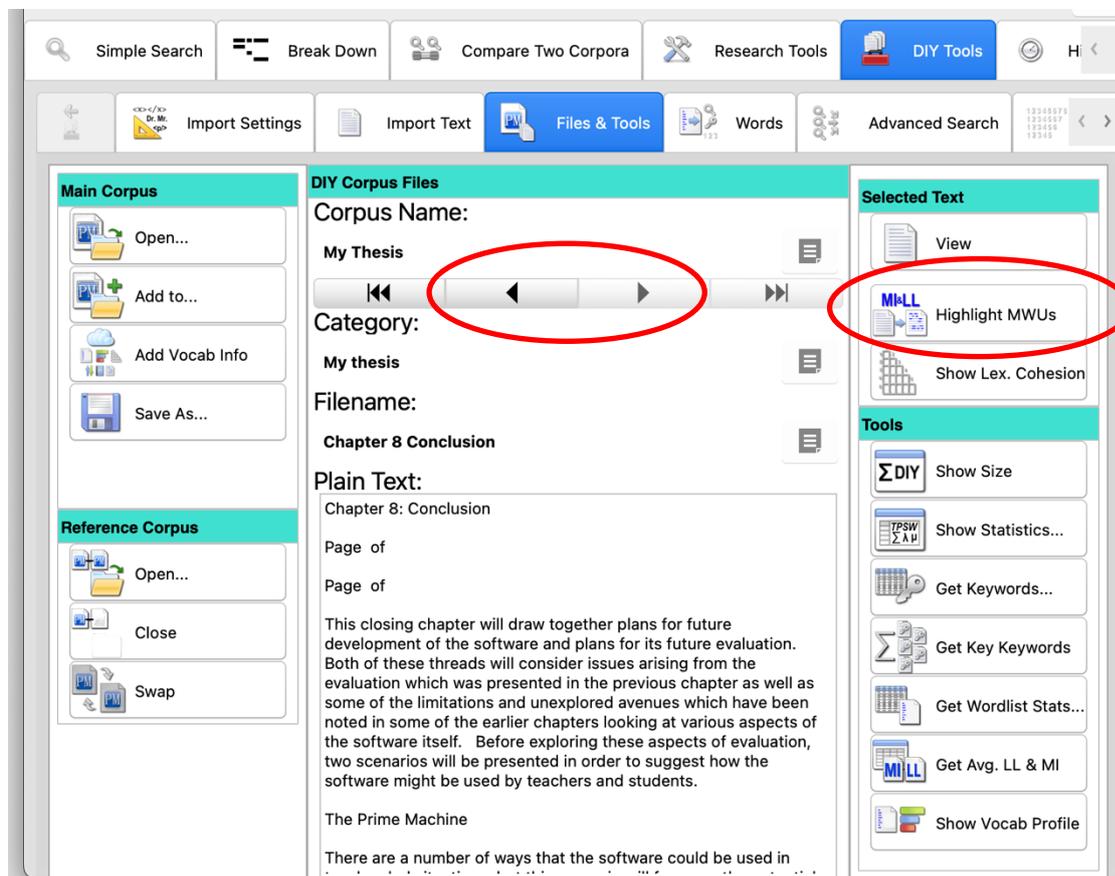
#	Score	LL Collocation	LL
1	10.00	number of	7256.33
2	8.00	number .. students	65.64
3	9.00	through .. medium	96.52
4	10.00	it .. important	1356.37
5	10.00	is .. important	300.95
6	10.00	becoming increasingly	694.99
7	6.00	becoming .. important	39.19
8	10.00	increasingly important	187.75
9	4.00	important .. understand	28.34
10	10.00	to understand	1313.39
11	7.00	language .. learning	45.98
12	6.00	learning strategies	38.87
13	10.00	as .. tool	167.26
14	1.00	a tool	17.05
15	8.00	tool for	70.17
16	7.00	their .. careers	48.56

The other way to get these scores is using one entire text from the DIY corpus. When you are working with chapters of a thesis as a DIY corpus, you probably don't want to send entire chapters to the server, so you could make a new smaller DIY corpus of just on paragraph. The "Paste from Clipboard" function is a quick way to make a small DIY corpus. You can set the name of the corpus before you paste text using the box above the button.

The screenshot shows the software's main interface. At the top, there are navigation tabs: Simple Search, Break Down, Compare Two Corpora, Research Tools, DIY Tools, and History. Below these are sub-tabs: Import Settings, Import Text, Files & Tools, Words, Advanced Search, and N-grams. The main area is divided into two columns. The left column has three sections: 'Import files' with 'Create new corpus...' and 'Add to corpus...' buttons; 'Import folder' with 'Create from folder...' and 'Add from folder...' buttons; and 'Paste from clipboard' with a text input field, a 'Make new corpus...' button (circled in red), and an 'Add to corpus...' button. The right column has a 'Drop Files' section with a dashed box containing a large downward arrow and the text 'TXT PDF DOCX DOC RTF PPTX PPT EPUB XML ZIP'.

To generate collocation highlighting results for a DIY corpus text, go to the Files & Tools tab under DIY Tools and find the "Highlight MWUs" button under Selected Text.

If you do want to send an entire text from a DIY corpus, you can navigate the texts on this tab by using the left and right arrows.



When you click the button, combinations of words will be sent to the server.

**Note: Using "Highlight MWUs" means chunks of your thesis will be transmitted over the internet. Efforts are made to make this secure, but you should be mindful of the risks of transmitting these kinds of data.**

The screenshot shows a software interface with a table of results. At the top, there are tabs for 'Table 1', 'Table 2', and 'Table 3', with 'Table 3' selected. Below the tabs is a title: 'MWUs in Chapter 2 Survey of Students and Teachers compared against BNC Academic (Incomplete Results)'. The table has columns: 'Line', 'Score', 'MI', 'DeltaP', 'LL', and 'Sentence'. A red box highlights a button labeled 'Break Down' in the top right corner of the table area. A callout box with a red border points to this button and contains the text: 'This button will copy the selected line to the Break Down tab.' The table content includes several rows of data, with the first row (Line 9) having a score of 6.56 and MI of 100. The sentences are excerpts from academic texts, such as 'As the number of students studying degree subjects through the medium of English grows, it is becoming increasingly important to understand the language needs and learning strategies of these students as they use English as a tool for their academic careers.'

The results break the text into sentences and each sentence has a set of collocation scores. Unlike the Break Down tab, the full stop at the end of each sentence is also included in the scores (so we get slightly different results for the first sentence).

### What does it mean?

When you are writing a dissertation, you will be making new connections between ideas and generating new knowledge. This means you will need to use words in new combinations – having some grey words in the results is a good thing! But your dissertation also needs to build on the ideas of others (properly cited and referenced of course). And because it is academic writing, you will need to express ideas using combinations of words common in academic texts. And because it is written in English, you will need to follow some of the conventions of combining words as used in the language more generally. The trick is to try to make sure the grey words are for the ideas you want to stand out, and for the most part everything else blends into green or light yellow. Experts reading your work will be struck by the less usual combinations of words, and the more conventional combinations will help carry the message along.

### What should I do with the highlighted combinations?

If we look at line 5 in the results above, we can see that the word *focussed* is in grey. This is actually because I have a preference for the less usual spelling of this word. It is correct, but *focused* with one *s* seems more common in the BNC: Academic. I can test this theory by changing the spelling and regenerating the results:

Line	Score	MI	DeltaP	LL	Sentence
1	6.65	100	100	50	some studies have focused specifically on this group

But in your work, you may be able to find some grey words that intrigue you – some combinations you thought were quite common and wouldn't need any attention have been coloured grey by this process.

Try to spot some of these and break down the phrases into different combinations, trying to see whether there are synonyms or other wordings which will make your expressions seem more natural. Don't take away everything that is creative. But if an expression is grey and it is describing a common idea, you probably should try to change it.

You can take pairs of words or look up words one by one. Or you can use the Break Down tab to help you select different patterns of words for comparison.

For this approach, you are looking at combinations of words up to four words apart. The computer doesn't know what you are trying to express, so it won't be able to solve any problems for you. You will need to think creatively about different synonyms you could try or different ways you could express the ideas.

Let's take an invented example and see how the software can help, and what we need to do for ourselves.

Line	Score	MI	DeltaP	LL	Sentence
1	5.68	<a href="#">100</a>	<a href="#">100</a>	<a href="#">40</a>	on another hand it is critical to employ statistical means

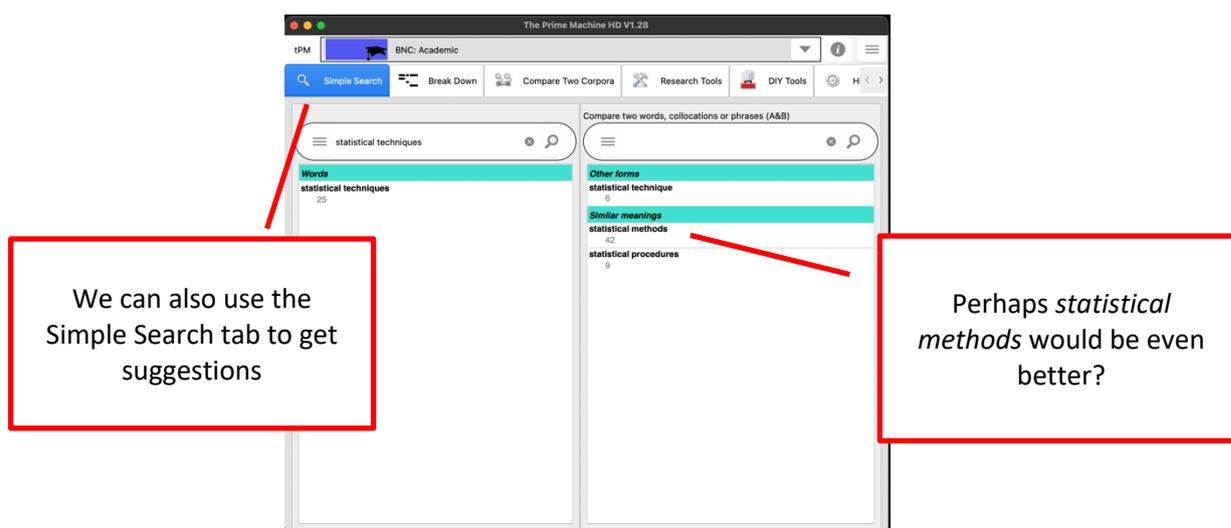
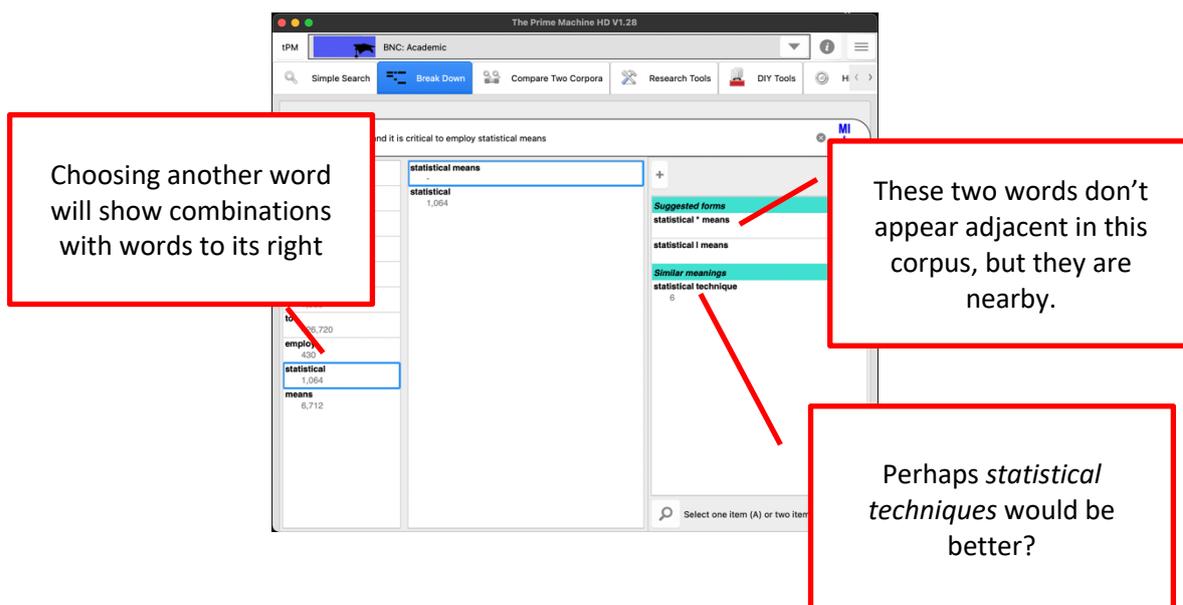
On the Break Down tab, we can click the hamburger or 蒸笼 button inside the search box and it will split the sentence into words, allowing us to look at combinations of up to five words starting at the currently selected item.

The first column shows the frequencies of each word in the sentence.

The second column shows combinations of words to the right of the selected word. If the combination is stored as a LL collocation it shows the number of hits.

If you click the + button, suggestions may appear.

You could check concordance lines if you couldn't guess "On the other hand" is better.



Looking at individual words and thinking of possible synonyms is also a good idea. We could explore the typical uses of *critical* vs. *important*, for example.

Once you have tried different combinations, you can see what the collocation score for the revised sentence would be.

MWUs in string compared against BNC Academic					
Line	Score	MI	DeltaP	LL	Sentence
1	7.01	100	100	100	on the other hand it is important to employ statistical methods

(Although the noun *means* does have meaning similar to *methods* in some contexts, perhaps the reason it seems strange in combination with *statistical* is because in statistics we also have another meaning of *means* – the averages of the data. Similarly, *critical* is probably a bad choice because it also has a special meaning in statistics – *critical values*.)

### 3.4 Reporting verbs (multiple concordance line searches)

#### What can I do?

In academic writing, we use a variety of verbs to introduce ideas from the sources we have used. These reporting verbs can communicate important information to the reader about your stance and the stance of the original writer. Your stance includes your attitude towards the piece of information you are reporting as well as the strength of your confidence or its applicability. Searching your own thesis for reporting verbs and looking at the reporting verbs in your sources can help you see whether you have used reporting verbs to effectively communicate these things.

#### How do I do it?

You could take each reporting verb and explore them one by one using the Wordlist Tools search box. But tPM also has a way of combining the results of several words and/or phrases using the Advanced Search tab.

Make sure you use the Advanced Search tab on the DIY Tools menu, rather than the tab under Research Tools!

Copy or type into the Node word(s) search box a list of reporting verbs. Remember to include different verb forms (present tense, past tense, etc.).

**argues states claims suggests mentions reports finds explains argued stated claimed suggested mentioned reported found explained**

To get a quick overview, you can use the Get Frequencies button.

Click the Get Lines button if you only want to search inside your thesis; click the Compare button to show results for the reference corpus too. See page 8 for information about using a readymade reference corpus or a second DIY corpus (e.g. of your sources).

The screenshot shows the 'Advanced Search' tab in the software. The 'Node word(s)' field contains 'crucial crucially'. The 'Filters' panel on the right has 'Use \* wildcards' checked. The 'Multiple Search Tools' section has 'Get Frequencies' and 'Get Lines' buttons circled in red. Three red callout boxes provide instructions: one points to the plus button in the search field, another points to the search field itself, and a third points to the 'Use \* wildcards' checkbox.

If you only enter one word, the plus button here will retrieve different word forms for the word, using matches in the readymade corpus

On this tab, put spaces between words for a OR search.

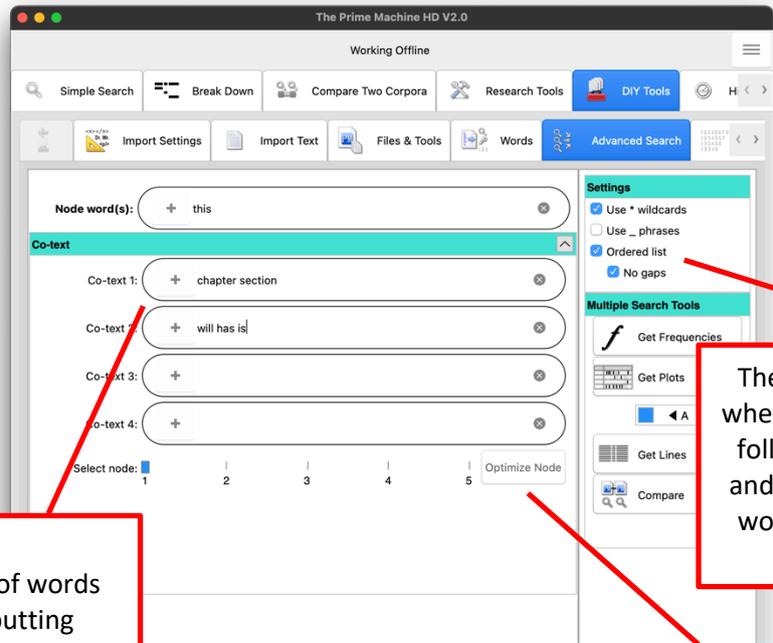
You can try to use wildcard searches like `argu*` but sometimes this finds too many matches.

For searches using more than one word, there are two approaches. If you wanted to include results for *according to* in these results, it would mean there would be a mixture of single words and multiple words in the query, so the best option is "use \_ phrases", where you simply put the \_ between the words with no spaces.

### argues states claim according\_to

The alternative approach is to keep "use \_ phrases" unticked and to use the Co-text boxes. Start with the top Node(s) box and work downwards one box at a time to represent each slot. For example, *according to* would have *according* in the top Node(s) box and *to* in the first Co-text box. If you wanted to look for *this study showed* and combine it with *this research showed* and *this experiment showed*, you would enter *this* in the Node(s) box, *study research experiment* (separated by spaces) first Co-text box and *showed* in the second Co-text box.

If you do not want to find hits where there are additional words in-between, you need to tick the "no gaps" option. If you want to include results where words are in different orders, you can untick the "ordered list" option.



The screenshot shows the 'Advanced Search' interface of 'The Prime Machine HD V2.0'. The 'Node word(s):' field contains 'this'. Below it are four 'Co-text' boxes: 'chapter section', 'will has is', and two empty boxes. The 'Settings' panel on the right has 'Use \* wildcards' checked, 'Use \_ phrases' unchecked, 'Ordered list' checked, and 'No gaps' checked. The 'Multiple Search Tools' panel includes 'Get Frequencies', 'Get Plots', 'Get Lines', and 'Compare'. An 'Optimize Node' button is at the bottom right. Three red callout boxes provide instructions: one points to the co-text boxes, one points to the 'No gaps' setting, and one points to the 'Optimize Node' button.

Enter strings of words vertically, putting options for each slot in a different box.

These settings control whether the words must follow the same order and whether additional words in-between are allowed.

For readymade corpus searches, the slot containing the fewest hits will be the easier for the server to use as the node. If this button is enabled when you click Compare, you can automatically select the optimal node.

If you use a readymade corpus for the search, it is possible that the Node(s) slot contains a very high frequency item such as *this* or *the*. It is much more efficient for the server to use a lower frequency word as the node because it will find all the hits of the node first and then filter out results which do not match the pattern. The Optimize Node button will light up and you can click it to adjust which search box is used as the node. Or you can ignore it and just click Compare a second time... but this could mean a delay in getting the results.

### What does it mean?

It is not surprising to find that you use a small set of reporting verbs. In my own thesis, we can see I've used *explained*, *reported* and *found* rather more often than *argued* or *suggested*. This matches my reasons for using different sources – I wanted to take ideas from sources about what was important, what should be considered and the results of previous studies to help inform my own design. However, a frequency table like the one below does help you spot any rarer choices which could stand out because the choice is different.

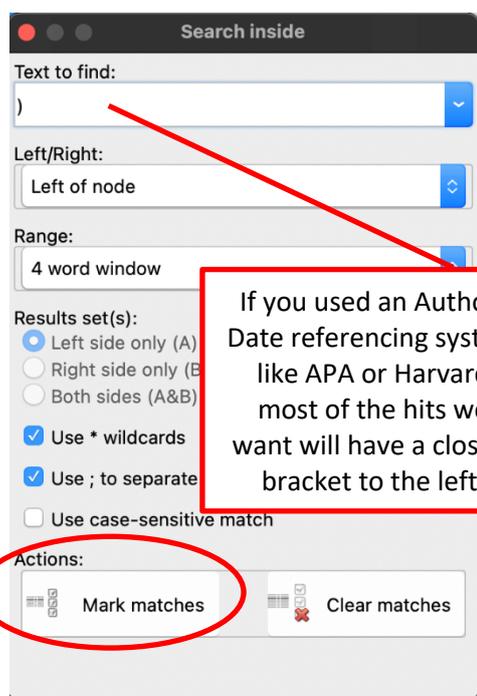
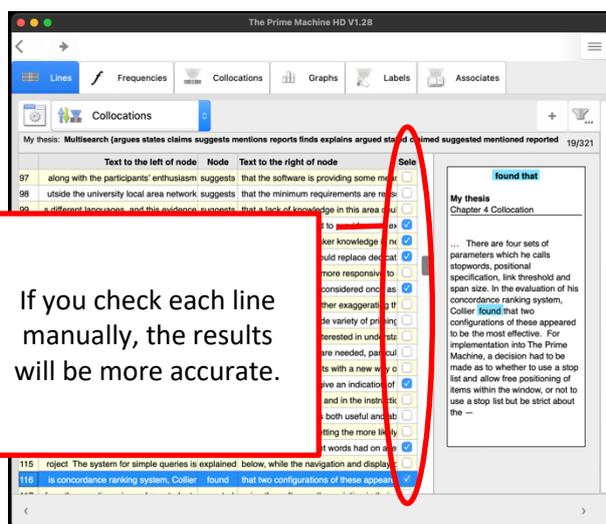
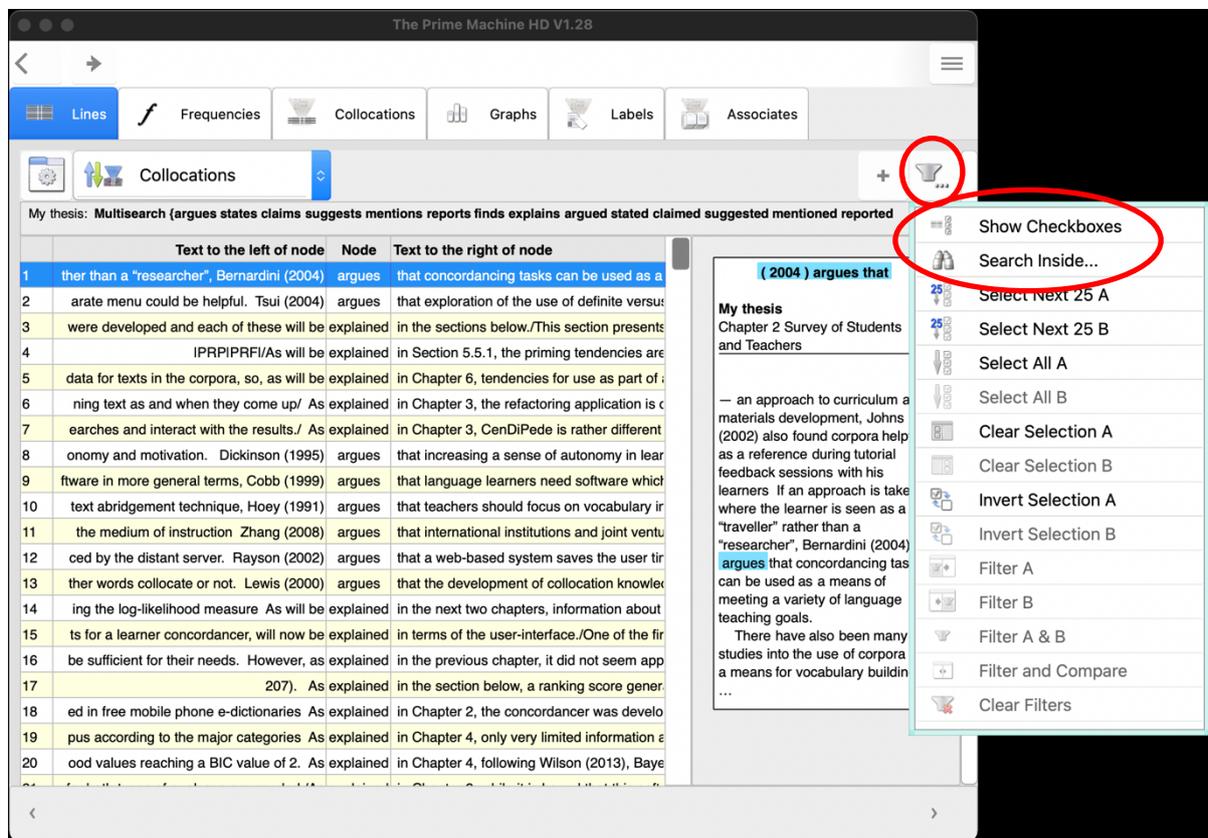
	Pattern	Frequency	Per Thousand
1	explained	61	0.53
2	reported	53	0.46
3	found	41	0.35
4	mentioned	35	0.30
5	argued	27	0.23
6	argues	22	0.19
7	suggests	21	0.18
8	explains	15	0.13
9	suggested	14	0.12
10	claims	9	0.08
11	reports	7	0.06
12	states	7	0.06
13	claimed	4	0.03
14	stated	4	0.03
15	mentions	1	0.01

Looking at concordance lines can also help you look at different choices and the corresponding authors - look to the left for authors and to the right for patterns of usage. It also shows that many of my *explained* hits are actually referring to other parts of my own thesis, rather than introducing the research of others. For *argues* it is worth looking several words to the right to see whether different modal verbs match the original author's and/or your own stance. The concordance card is sometimes needed to look so far after the node. The red lines in the figure below have been added manually for demonstration purposes.

The screenshot shows the 'The Prime Machine HD V1.28' interface. The search bar contains 'My thesis: Multisearch {argues states claims suggests mentions reports finds explains argued stated claimed suggested mentioned reported 321}'. The concordance table below shows results for the node 'argues'. The selected node is '( 2004 ) argues that', which is highlighted in a concordance card on the right. The card shows the full context of the sentence: '— an approach to curriculum and materials development, Johns (2002) also found corpora helpful as a reference during tutorial feedback sessions with his learners. If an approach is taken where the learner is seen as a "traveller" rather than a "researcher", Bernardini (2004) argues that concordancing tasks can be used as a means of meeting a variety of language teaching goals. There have also been many studies into the use of corpora as a means for vocabulary building. ...'.

	Text to the left of node	Node	Text to the right of node
1	ther than a "researcher". Bernardini (2004)	argues	that concordancing tasks can be used as a
2	arate menu could be helpful. Tsui (2004)	argues	that exploration of the use of definite versus
3	were developed and each of these will be	explained	in the sections below./This section presents
4	IPRPIPRFI/As will be	explained	in Section 5.5.1, the priming tendencies are
5	data for texts in the corpora, so, as will be	explained	in Chapter 6, tendencies for use as part of
6	ning text as and when they come up/ As	explained	in Chapter 3, the refactoring application is c
7	earches and interact with the results./ As	explained	in Chapter 3, CenDiPede is rather different
8	onomy and motivation. Dickinson (1995)	argues	that increasing a sense of autonomy in lear
9	ftware in more general terms, Cobb (1999)	argues	that language learners need software which
10	text abridgement technique, Hoey (1991)	argues	that teachers should focus on vocabulary ir
11	the medium of instruction Zhang (2008)	argues	that international institutions and joint ventu
12	ced by the distant server. Rayson (2002)	argues	that a web-based system saves the user tir
13	ther words collocate or not. Lewis (2000)	argues	that the development of collocation knowle
14	ing the log-likelihood measure As will be	explained	in the next two chapters, information about
15	ts for a learner concordancer, will now be	explained	in terms of the user-interface./One of the fir
16	be sufficient for their needs. However, as	explained	in the previous chapter, it did not seem app
17	207). As	explained	in the section below, a ranking score gener
18	ed in free mobile phone e-dictionaries As	explained	in Chapter 2, the concordancer was develo
19	pus according to the major categories As	explained	in Chapter 4, only very limited information e
20	ood values reaching a BIC value of 2. As	explained	in Chapter 4, following Wilson (2013), Baye

As noted earlier, some of the hits for these reporting verbs in my thesis were not reporting cited literature, but were referring to other parts of my thesis. If you want to look more closely at only the hits where they are really reporting verbs with citations, you can use the filter menu. You could select lines manually by clicking "Show checkboxes" and going through line by line. Or you could try using the "Search inside..." function.



If you used an Author-Date referencing system like APA or Harvard most of the hits we want will have a closing bracket to the left.

After selecting the lines using one of the methods, go back to the Filter menu and choose Filter A. The results will now only contain the ones matching our criteria. As before, the red lines here have been added manually to demonstrate where you could be looking as you explore patterns of use and modality.

	Text to the left of node	Node	Text to the right of node
1	ther than a "researcher", Bernardini (2004)	argues	that concordancing tasks <b>can be used as a</b>
2	arate menu could be helpful. Tsui (2004)	argues	that exploration of the use of definite versus
3	onomy and motivation. Dickinson (1995)	argues	that increasing a sense of autonomy in lear
4	ftware in more general terms, Cobb (1999)	argues	that language learners <b>need</b> software which
5	text abridgement technique, Hoey (1991)	argues	that teachers <b>should focus</b> on vocabulary ir
6	the medium of instruction. Zhang (2008)	argues	that international institutions and joint ventu
7	ced by the distant server. Rayson (2002)	argues	that a web-based system <b>saves</b> the user tir
8	ther words collocate or not. Lewis (2000)	argues	that the development of collocation knowle
9	ddition to these issues, as Anthony (2004)	argues	as he presents his classroom concordance
10	n by Wermter and Hahn (2006) where it is	argued	that compared to the other statistical exten
11	oey's theory of Lexical Priming (2005), it is	argued	that some words <b>are actually primed</b> to occ
12	y cannot. As Frankenberg-Garcia (2011)	argues	, an important aspect of the use of referenc
13	l. In the early days of DDL, Johns (1986)	explains	that the computer <b>would take</b> about 50 sec
14	rs on Data-Driven Learning, Johns (1991)	explains	that students <b>often</b> come to concordancers
15	tionaries. Over 30 years ago, Nesi (1987)	argued	that dictionaries <b>need</b> to provide more exan
16	he classroom. However, Stevens (1995)	suggests	that near native speaker knowledge is nee
17	n dedicated software tools. Meyer (2002)	suggests	that shared scripts <b>could</b> replace dedicati
18	t double counting of items, Sinclair (1991)	argued	that each token was considered once as nc
19	erent levels of word frequency, Hai (2008)	found	that the most frequent words had on averag
20	he study by O'Donnell et al. (2012) which	found	that one in forty individual words showed a

### What should I do with the results?

Using a variety of reporting verbs merely for the sake of using a variety is not going to help make your ideas and your understanding clear. So don't try to replace reporting verbs unless you are sure their subtle difference in meaning matches your context.

Looking at the reporting verbs in a reference corpus can also help you see how published writers refer to research. You can read examples from the reference corpus to see:

- Are there patterns of grammar words to consider, and do these relate to meaning?
- Are there any theories or claims or results which need stronger or weaker signals?
- Do the subtle meanings of the reporting verbs match your intentions?
- Have you used past tense and present tense effectively?

Before changing your reporting verbs or trying out new ones, check the patterns of use in a reference corpus to make sure you understand how to use them and what they mean.

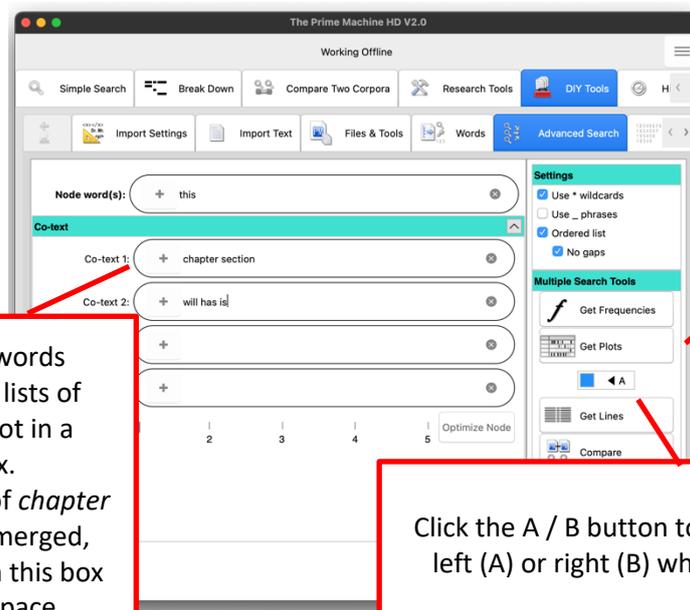
### 3.5 Tenses and modality (multiple concordance line searches)

#### What can I do?

One important aspect of your thesis is the way in which you signpost the reader, introducing what will be coming up in the next section and what has been accomplished in the previous section. Using tPM, you can search through your thesis for auxiliary verbs like *will* and *has* or metalanguage through words like *thesis*, *chapter* and *section*. You can check you haven't kept any signals of future intentions from an earlier draft of a proposal ("the participants will complete..." → "the participants completed...") and you can see the positions of *will* vs. *has* in a graphic representation using tPM's plots function.

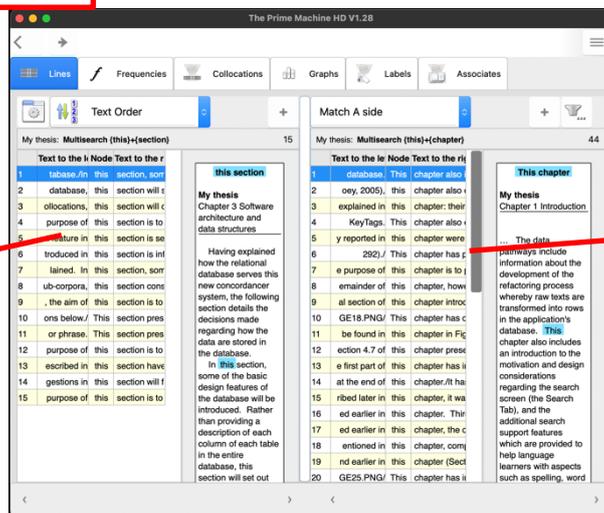
#### How do I do it?

If you want to compare your own thesis with a reference corpus, you can use the search box on the Wordlist tools tab to look up one word at a time (clicking the compare button to complete the search on both your own thesis and the reference corpus). If you want to compare *will* with *has* or *this chapter will* with *this chapter has*, you can use the Advanced Search tab. Perform one search first; then click the A/B button and do the second search.



Enter strings of words vertically, putting lists of words for each slot in a different box. If you want results of *chapter* and *section* to be merged, enter both words in this box separated by a space.

Click the A / B button to send the results to the left (A) or right (B) when you click Get Lines.



This side was produced with the A/B switch set on A (the default)

This side was produced with the A/B switch set on B (as above)

The results of plots will show each chapter (or text), with a vertical line to show the location of each hit in the text.

Text	Freq.	By Relative %
1 Chapter 1 Introduction	4	[Plot]
2 Chapter 2 Survey of Students and Teachers	2	[Plot]
3 Chapter 3 Software architecture and data structures	6	[Plot]
4 Chapter 4 Collocation	11	[Plot]
5 Chapter 5 Further features of Lexical Priming	14	[Plot]
6 Chapter 6 Metadata KeyTags and Key Associates	8	[Plot]
7 Chapter 7 Evaluation	9	[Plot]
8 Chapter 8 Conclusion	5	[Plot]
Total	59	[Plot]

**Plot Settings**  
Buckets: 10  
Plot column width: 20

**Plot Colours**  
Blues  
Greys  
Custom  
1. Powderblue  
2. Skyblue  
3. Dodgerblue  
4. Blue  
5. Navy  
6. Black

You can control the plot display by changing settings on the Plots tab of the Options screen. This can be accessed using the main hamburger or 蒸笼 menu.

Buckets are the number of divisions for the text. If we change it to 10, we will have one box for each 10% of running words. We can also make the plot lines wider if we wish.

Text	Freq.	By Relative %
1 Chapter 1 Introduction	4	[Plot]
2 Chapter 2 Survey of Students and Teachers	2	[Plot]
3 Chapter 3 Software architecture and data structures	4	[Plot]
4 Chapter 4 Collocation	10	[Plot]
5 Chapter 5 Further features of Lexical Priming	10	[Plot]
6 Chapter 6 Metadata KeyTags and Key Associates	3	[Plot]
7 Chapter 7 Evaluation	8	[Plot]
8 Chapter 8 Conclusion	3	[Plot]
Total	44	[Plot]

The plot above was produced using 10 buckets, so each square represents 10% of running words of each text. We can see that "this chapter" has a high density at the beginning of Chapter 7, and at the beginnings and ends of chapters overall.

### **What does it mean?**

There are different ways of writing the introductions and conclusions of chapters of a thesis. My own personal preference is to use *will* or *is to* in introductions (*this chapter will, the purpose of this chapter is to*), and then to use *has* or *was to* in the conclusions (*this chapter has shown, this chapter has described, the purpose of this chapter was to*). There are many ways to communicate the direction and progression of your thesis, but it is good to try to establish a consistent approach. Clearly, *will* and *has* will have a variety of functions in a text, but by looking for these words, you may be able to spot future intentions and retrospective reminders.

### **What should I do with the results?**

Ask yourself:

- Were the expectations raised by *will* fulfilled?
- Have I used *will* for something I wrote in a proposal and forgotten to change it for the final dissertation?
- Are the ways I have signalled what is coming up and what has just been discussed clear and consistent?
- Are my introductions and conclusions of chapters or sections overly formulaic? (I am guilty of this too, but it is something to reflect on!)

## **3.6 Corpus statistics and other wordlists (average word, sentence and paragraph lengths; measures of language complexity; vocabulary profiling)**

### **What does it do?**

You can generate a range of different kinds of statistics to get insights into the language of your thesis. You can see the average lengths of words, sentences, and paragraphs in each of your chapters and in the thesis as a whole. You can also see statistics based on the range of different vocabulary. For example, you can see the Type-Token Ratio (different word forms divided by total running words) and the percentage of words matching the Academic Word List.

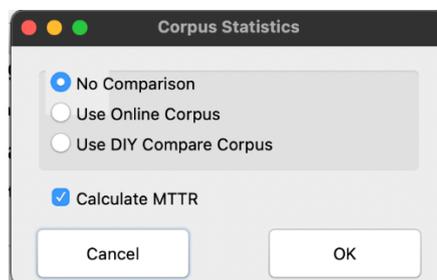
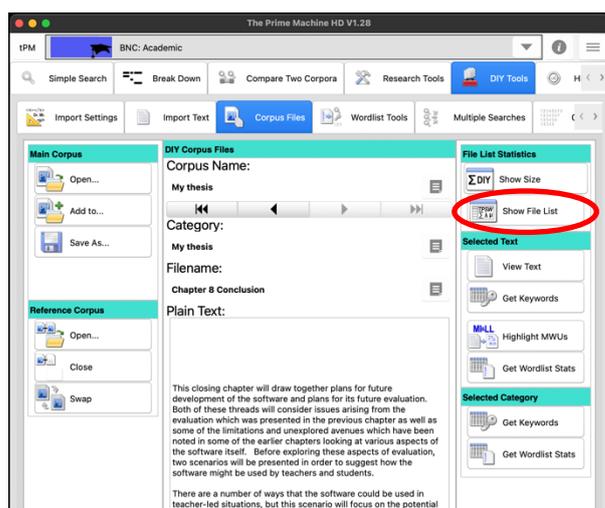
### **How do I do it?**

As you may want to compare the statistics for your thesis against the baseline of a reference corpus, you should think about which reference corpus to use first. For readymade corpora (from the online server) you could use BNC: Academic or a Hindawi corpus for a specialist academic discipline. If you want to use the DIY corpus of your sources as a reference corpus, you need to have loaded that first. You can save and load DIY corpora from the Files & Tools tab.

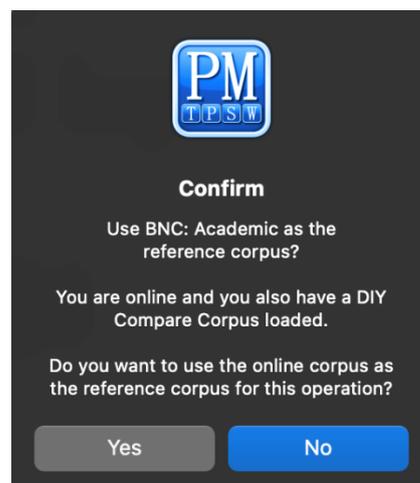
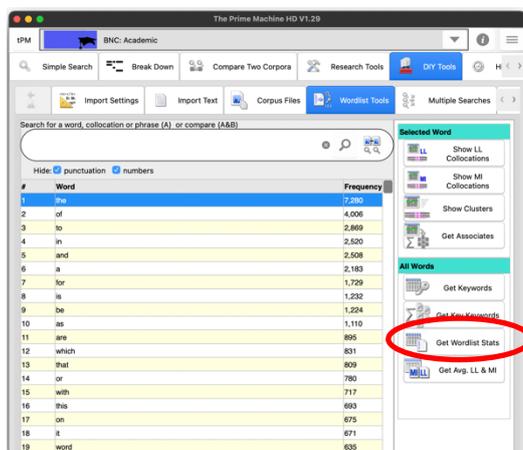
To view statistics about the average lengths of words, sentences and paragraphs, and several other statistics related Type-Token Ratio, simply click the Show File List button on the Files & Tools tab under DIY Tools.

You will be prompted to decide whether you want the statistics to show a comparison with a readymade Online Corpus or a second DIY corpus (if you have loaded one). The option for MTTR (Moving Type Token Ratio) only takes a few additional seconds.

Note: MTTR is based on a method called MATTR – see the tPM Help Selected Bibliography available from <https://www.theprimemachine.net/help.html> for a link.



To get a summary of matches on various wordlists (including the Academic Word List), you can click the Get Wordlist Stats button on the Wordlist Tools tab. It is also possible to get these statistics for your DIY corpus on a text by text basis (i.e. one chapter at a time) or on a category basis and those buttons are on the Corpus File tab under Selected Text and Selected Category respectively.



If you have a second DIY corpus loaded as a reference corpus, you will be prompted to confirm whether you want to use the online readymade corpus (e.g. BNC: Academic) or your second DIY corpus as a reference.

For the Wordlist Stats function, you need to be connected to the tPM server as the wordlist from your DIY corpus will be sent to the server to check each item against the various wordlists held there.

If you use a second DIY corpus as a reference corpus, two tables of results will be generated: one table with your thesis compared against the baselines of the second DIY corpus, and one table with the second DIY corpus compared against the baselines of your thesis.

### What does it mean?

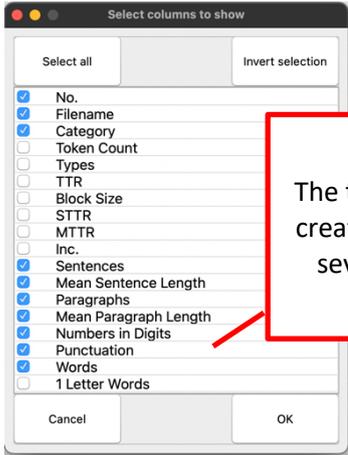
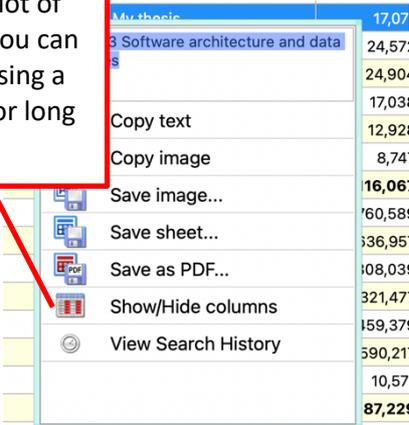
The results from Show File List are displayed on a very wide table.

No.	Filename	Category	Token Count	Types	TTR	Block Size	STTR	MTTR	Inc.	S
1	Chapter 1 Introduction	My thesis	1,684	511	0.30	400	0.46	0.46	1	
2	Chapter 2 Survey of Students and Teachers	My thesis	9,123	1,592	0.17	400	0.49	0.49	1	
3	Chapter 3 Software architecture and data structures	My thesis	17,071	2,345	0.14	400	0.50	0.50	1	
4	Chapter 4 Collocation	My thesis	24,572	2,870	0.12	400	0.49	0.49	1	
5	Chapter 5 Further features of Lexical Priming	My thesis	24,904	2,952	0.12	400	0.48	0.48	1	
6	Chapter 6 Metadata KeyTags and Key Associates	My thesis	17,038	2,439	0.14	400	0.48	0.48	1	
7	Chapter 7 Evaluation	My thesis	12,928	1,886	0.15	400	0.47	0.48	1	
8	Chapter 8 Conclusion	My thesis	8,747	1,473	0.17	400	0.49	0.49	1	
	<b>Total</b>		<b>116,067</b>	<b>6,744</b>	<b>0.06</b>	<b>400</b>	<b>0.49</b>	<b>0.48</b>	<b>8</b>	
10		Humanities and Arts	3,760,589	70,019	0.02	400	0.50	0.50	355	
11		Medicine	1,636,957	32,100	0.02	400	0.47	0.47	300	
12		Natural Science	1,308,039	25,000	0.02	400	0.47	0.47	300	
13		Politics, Law and Education	5,321,477	100,000	0.02	400	0.47	0.47	300	
14		Social Science	5,459,379	100,000	0.02	400	0.47	0.47	300	
15		Technology and Engineering	590,217	10,000	0.02	400	0.47	0.47	300	
16		Other	10,571	2,402	0.23	400	0.50	0.50	4	
17	<b>Online Corpus Total</b>	<b>BNC: Academic</b>	<b>18,087,229</b>	<b>150,689</b>	<b>0.01</b>	<b>400</b>	<b>0.48</b>	<b>0.48</b>	<b>2,244</b>	

The total token count includes punctuation!

TTR is affected by the length of the text, so STTR and MTTR are more reliable.

There are a lot of columns, so you can hide some using a double-click or long tap.



The table below was created by unticking several columns.

No.	Filename	Category	Sentences	Mean Ser	Paragrap	Mean Para	Numbers	Punctuatic	Words	Mean Word
1	Chapter 1 Introduction	My thesis	48	35.08	16	105.25	13	111	1,560	5.14
2	Chapter 2 Survey of Stude	My thesis	277	32.94	56	162.91	153	1,000	7,970	5.18
3	Chapter 3 Software archit	My thesis	534	31.97	138	123.70	179	1,892	15,000	4.99
4	Chapter 4 Collocation	My thesis	1,112	22.10	532	46.19	590	2,816	21,166	4.96
5	Chapter 5 Further features:	My thesis	2,218	11.23	1,737	14.34	1,271	3,139	20,494	4.92
6	Chapter 6 Metadata KeyTe	My thesis	1,103	15.45	737	23.12	562	1,910	14,566	4.87
7	Chapter 7 Evaluation	My thesis	525	24.62	228	56.70	265	1,321	11,342	4.94
8	Chapter 8 Conclusion	My thesis	264	33.13	70	124.96	46	713	7,988	5.08
	<b>Total</b>		<b>6,081</b>	<b>19.09</b>	<b>3,514</b>	<b>33.03</b>	<b>3,079</b>	<b>12,902</b>	<b>100,086</b>	<b>4.97</b>
10		Humanities and Arts	127,480	29.50	21,705	173.26	33,622	451,350	3,275,617	4.89
11		Medicine	66,852	24.49	14,195	115.32	34,811	185,974	1,416,172	5.37
12		Natural Science	55,957	23.38	11,129	117.53	30,244	165,256	1,112,539	5.04
13		Politics, Law and Education	190,531	27.93	42,839	124.22	80,849			
14		Social Science	237,917	22.95	42,469	128.55	83,450			
15		Technology and Engineering	29,878	19.75	6,698	88.12	9,898			
16		Other	563	18.78	151	70.01	358			
17	<b>Online Corpus Total</b>	<b>BNC: Academic</b>	<b>709,178</b>	<b>25.50</b>	<b>139,186</b>	<b>129.95</b>	<b>273,232</b>	<b>2,127,963</b>	<b>15,686,034</b>	<b>5.01</b>

You may need to resize some columns

The results of the Wordlist Statistics are sorted using a log-likelihood statistic, showing how the matches in your thesis stand out from the results of the reference corpus (in this example the BNC: Academic).

	Wordlist	Study Freq.	Study Per Thousand	Ref. Freq.	Ref. Per Thousand	Arrows	LL Bayes
1	Academic Word List	10,244	88.26	1,367,114	75.58	↑	232.39 Very strong evidence in favour
2	General Service List 2	5,424	46.73	750,471	41.49	↑	73.27 Very strong evidence in favour
3	Modals	1,705	14.69	221,854	12.27	↑	51.91 Very strong evidence in favour
4	Modals Subgroup 2	873	7.52	109,646	6.06	↑	37.59 Very strong evidence in favour
5	Modals Subgroup 1	576	4.96	69,024	3.82	↑	36.23 Very strong evidence in favour
6	1st & 2nd Pers. Pronouns	117	1.01	101,568	5.62	≈ 5x ↓	
7	General Service List 1	73,319	631.70	11,449,894	633.04	↓	
8	Archaic Pronouns	0	0.00	426	0.02	↓	
9	Punctuation	5,879	50.65	1,060,127	58.61	↓	
10	Modals Subgroup 3	256	2.21	42,901	2.37	↓	
11	Personal Pronouns	1,602	13.80	481,408	26.62	↓	
12	Function Words	47,625	410.32	7,627,311	421.70	↓	
13	Positive words	1,224	10.55	244,749	13.53	↓	
14	Negative words	650	5.60	336,901	18.63	≈ 3x ↓	

My thesis seems to use modals more frequency than general academic texts.

TTR is calculated by taking all the different words (types) and dividing by the number of running words (tokens). Texts where there is a high repetition of the same words will have a TTR closer to zero. Texts with a wider variety of words will be closer to 1. However, TTR is influenced by the overall text length, so it is more meaningful to look at blocks of text and then take the average of those blocks. STTR breaks each chapter of your thesis into blocks of 400 words and measures the TTR for each block and then takes the average of those blocks. MTTR is similar, but it uses many more blocks, as the blocks are a moving window of 400 words. MTTR will be more accurate if there are sections of the text which are less rich than others. However, in many cases you will find STTR and MTTR are very similar. Typically, we would expect formal academic writing (like a thesis) to have a STTR and MTTR of just under 0.5. We can see in the table on the previous page that most chapters of my thesis have a STTR matching that figure. Perhaps Chapter 1 and Chapter 7 are a little lower. Chapter 1 was my shortest chapter and had the usual repetitive language about the structure of the thesis. Chapter 7 will have had several graphs and tables, and this may have affected the results too.

The second table on the previous page shows statistics based on average sentence length, paragraph length and word length. The figures for sentences and paragraphs will depend on how well tPM has been able to recognise sentences and paragraph breaks. If you have loaded DOCX files (as in my example) it is possible that some breaks counted in tPM as paragraphs will actually be bullet points, captions, and data from tables. It would be possible for you to edit out tables and captions on a copy of your thesis before you load it. Or you can bear these limitations in mind while you review these data.

Looking at the first three chapters of my thesis – chapters with fewer graphs and tables – we can see that my average sentence length is rather longer than those in published academic texts. My paragraph lengths are similar to the reference corpus. Word lengths are also fairly similar.

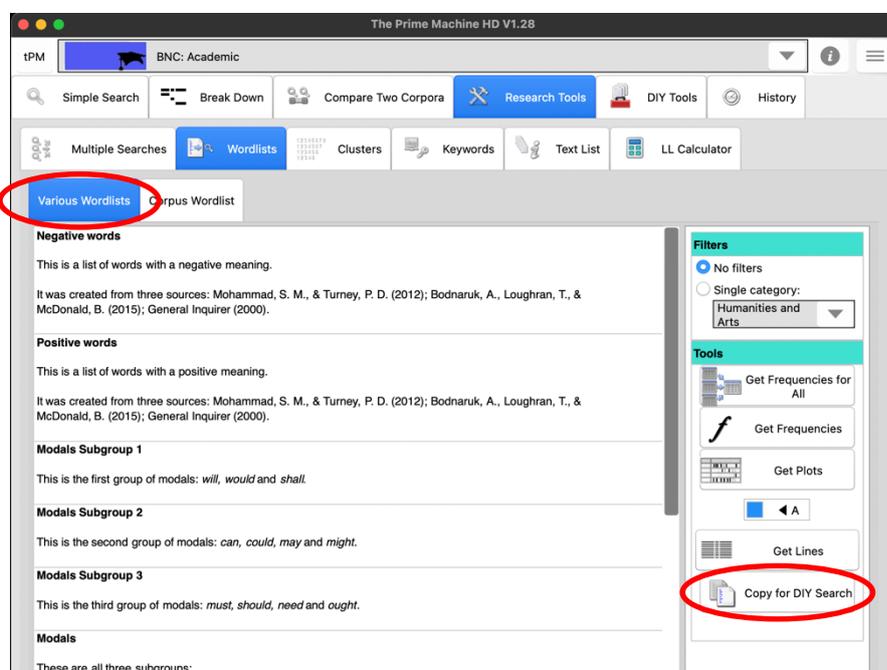
If you have longer sentences (like me), you might consider trying to break some sentences down to help the reader follow your message. If you have shorter sentences, it could be that you could try to combine more ideas or add details to express ideas more precisely.

The Wordlist Statistics can help you see to what extent your thesis matches or contrasts with the baselines of the reference corpus. Typical features of academic texts would include higher proportions from the Academic Word List (but typically only between 5% and 10%) and lower proportions of first and second person pronouns.

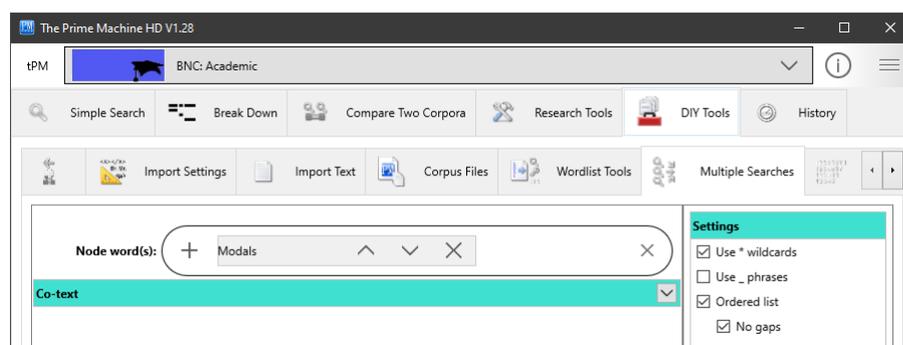
### What should I do with the results?

The statistics are just statistics... they can't tell you if it is good writing and you don't need to try to make your thesis match the baselines of your reference corpus. However, they may give you some useful starting points to help you when you re-read your thesis closely to revise it, or to discuss with your supervisor.

To find out more about these wordlists, you can find information on the Wordlists tab under Research Tools. It is also possible to use some of these shorter wordlists in DIY corpus searches. Some wordlists are too long to be used in this way, however.



After clicking the copy button, you'll be taken to the Advanced Search tab under DIY Tools and you will see a special box appear inside the Node word(s) search box. You can move this up and down to other slots, so you can also use these shorter wordlists in combination with other words (e.g. *this + modal*).



For some of the background to these methods and approaches, please see the tPM Help Selected Bibliography available from <https://www.theprimemachine.net/help.html>.



Dr. Stephen Jeaco - 杰大海  
[www.theprimemachine.net](http://www.theprimemachine.net)



This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/> or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA. If you use this document, please include a link to the tPM Help Selected Bibliography available from <https://www.theprimemachine.net/help.html>.

First published: Thursday, 17 March 2022

Last updated: Sunday, 10 September 2023