Research Tools: DIY Text Tools

As with the other Research Tools, the DIY Text Tools are primarily designed for small research projects at the undergraduate level.

What are the DIY Text Tools for?

These tools are designed to help you analyse small amounts of text (hundreds of thousands of words, rather than millions or billions of words). With these tools you can:

- Import plain text
- Generate wordlists and concordance cards and lines for your own texts;
- Calculate collocations and display clusters for your own texts;
- Use *The Prime Machine*'s online corpora as a reference with which to compare the relative frequencies of words (Wordlist statistics, Key Words, Key Key Words and Key Associates);
- Use an online corpus to determine the average mutual information score for a text or a small collection of texts by comparing pairs of words against an online corpus.
- Compare two sets of DIY Corpora with each other.

Using *The Prime Machine* to process and explore your own DIY corpus has some advantages but there are also a number of limitations. One advantage is being able to use any of the preloaded online corpora as a reference corpus. Also, provided your texts are in plain text (see below) using tPM for DIY corpus analysis is relatively simple. However, limitations include slow processing speed for larger datasets and a lack of fine-tuning for the calculations you perform. There are some limits on the number of different words (types), the corpus size (tokens) and the number of texts for some of the online operations. Users connected to tPM's Home Network will be able to process larger collections of texts than users who connect from offsite. Compared with other leader corpus tools, the scope for fine-tuning the collocation and key word measures is also rather limited. At the time of writing, DIY Tools are a new feature of *The Prime Machine* and additional features or fine-tuning of existing features may be made available in the future.

If you want to perform some kind of analysis of text in your DIY Corpus and the current version of The Prime Machine *is too limited, do consider contacting the developer and making a feature request. The link to feedback can be found on tPM's website:* www.theprimemachine.com

Getting started with DIY Tools

You can find the DIY Text Tools by going to the main Search Tab, choosing Research tools and then choosing DIY Text Tools.
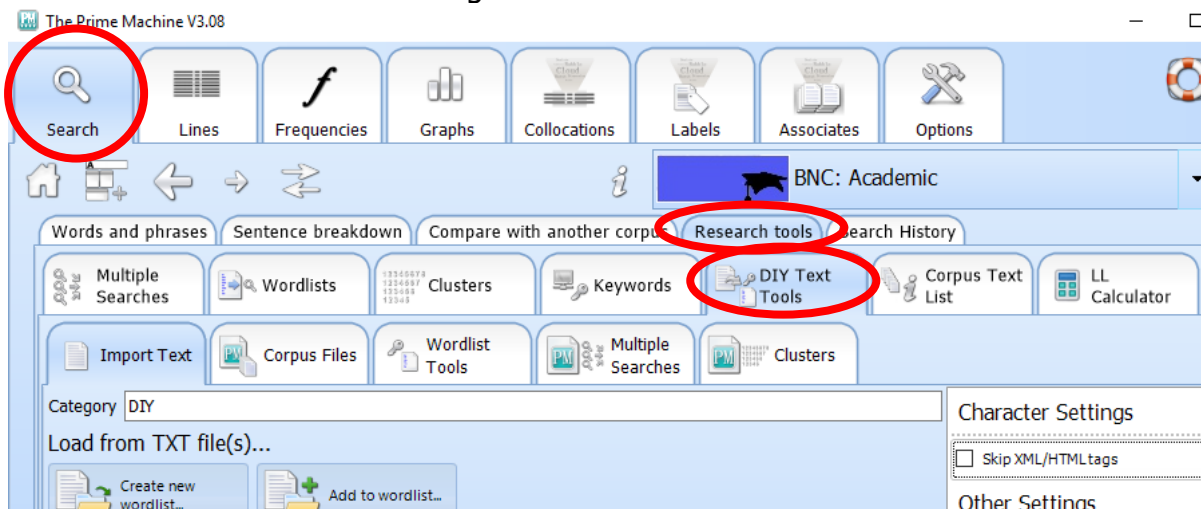
- You can access the DIY Text Tools when you are working offline. You can import text, view concordance lines and collocations while working offline. Also, if you load another DIY corpus as a reference corpus, you can get Key Words, Key Key Words and Key Associates while offline.

Importing Text

Before you can work with your own texts in *The Prime Machine*, you need to ensure they are in a file format which can be read by the tPM system.

The files should:
- Be in plain text format (.TXT);
- Unicode (UTF-16), UTF-8 or your system default encoding.

Files may:
- Include XML or HTML tags (e.g. <B>, </p>). You can choose to ignore these or to use a small set to indicate sentence or paragraph breaks;

- If your files are Word Documents, PDF, or other formats, you will first need to convert them to plain text, or copy and paste the text into the import text box.
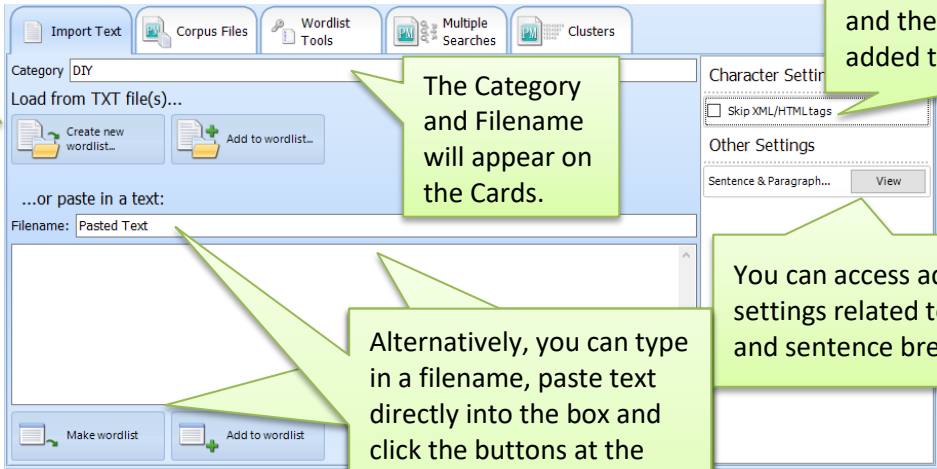

- File encoding is very important. Special characters and even simple punctuation such as ' and " may appear incorrectly if you do not use the correct encoding when saving your TXT file. Always check the Wordlist and look at some sample concordance lines for your corpus to make sure everything looks correct. If there are strange symbols, strange Chinese characters or unwanted spaces after ', check the file encoding and import the text again.

- Importing several texts into a new corpus is faster than importing texts one by one. If you put all the texts into a folder, you can easily select them all and import them at the same time.

The page for importing text looks like this:

If your text file includes tags like <b> <h1>, etc., tick this box to skip over these tags and they won't be added to the corpus.

Use these buttons to load text from files.

The Category and Filename will appear on the Cards.

You can access advanced settings related to paragraphing and sentence breaks here.

Alternatively, you can type in a filename, paste text directly into the box and click the buttons at the bottom to create a new corpus or add to the current corpus.

To import text from a file (or a set of files):
1. Type in a Category for the texts you want to import. You can change this text-by-text later (see note on the Corpus Files page below).
2. Check the files to see whether they contain XML or HTML tags. If so, tick the "Skip XML/HTML tags" box on the right.
3. Click "Create new wordlist…"
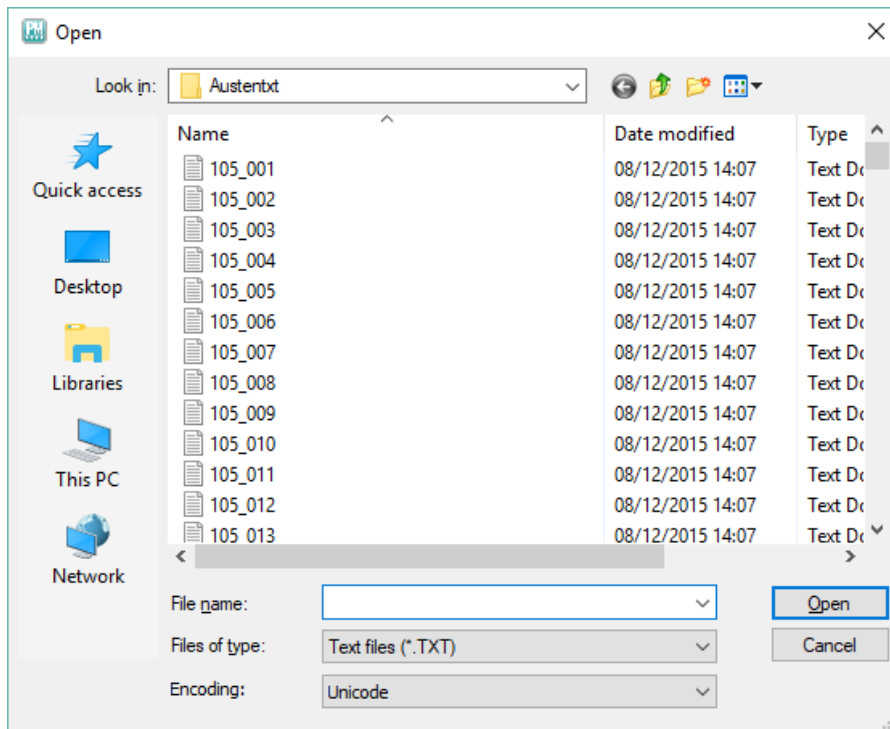
4. Find the text file (or files). As with other Windows programs, to select more than one file, choose the first file and then hold SHIFT and choose the last file in a series (or hold CTRL and choose each of the other files you want).
5. Check the encoding, using the drop-down list to select Unicode (UTF-16), UTF-8 or System default.
6. Click Open.
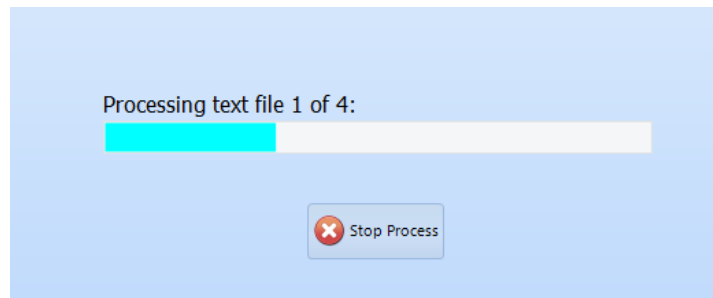
A screen will show progress as each file is opened and then a wordlist is generated, along with a list and counts for pairs of words. You can cancel the process by clicking the Stop Process button.

When the import process is complete, the Wordlist Tools tab will be displayed.

To import text using copy and paste:
1. Type in a Category and a Filename for the text you want to import. You can change these later (see note on the Corpus Files page below).
2. Paste the text into the box.
3. Click **Make wordlist** to create a new corpus containing just this text, or click **Add to wordlist** to add this text to the current corpus.

Simple Wordlists

A simple Wordlist can be found for the current corpus on the Wordlist Tools page. At the top of the list you will see the number of words (=tokens) and types (different words). These totals will include punctuation, but you can choose to show or hide punctuation in the wordlist by using the check box.

You can hide or show punctuation in the list, but punctuation will always be counted in the totals

DIY Wordlist has 11,269 words and 1,962 types
☑ Hide punctuation

#	Word	Frequency
1	the	388
2	of	367
3	and	324
4	to	278
5	a	208
6	in	178
7	had	170
8	her	125
9	was	125
10	she	112
11	it	106

The wordlist is sorted in descending order by frequency (the most frequent words will be at the top).

Like other tables in *The Prime Machine*, to copy or save the list of results as a picture or spreadsheet simply right-click or double-click on the table.

Loading and Saving Corpus Files

After you have created your DIY corpus, you should save it in a special tPM DIY Corpus file format so you can load it more quickly in the future. You can do this on the Corpus Files page.

A previously saved DIY corpus can be reopened here.

Remember to save your DIY Corpus

Load/Save DIY Corpus Files

Open DIY Corpus...   Add to DIY Corpus...   Save DIY Corpus...   Open DIY Corpus for Comparing...   Close DIY Compare Corpus   Show List of Files

<< 　✏ **DIY** ✏ 105_005 　　　>>

You can navigate through the texts in your corpus and edit the Category or Filename here.

Each original text file will be displayed here.

Get Keywords for Current Text   Get Keywords for Current Category   Get Wordlist Statistics for Current Text   Get Wordlist Statistics for Current Category

File list

After you have imported texts or reopened a DIY corpus, you can view a list of the filenames and the number of words for each file by clicking on the Show List of Files button which can be found on the Corpus Files page.
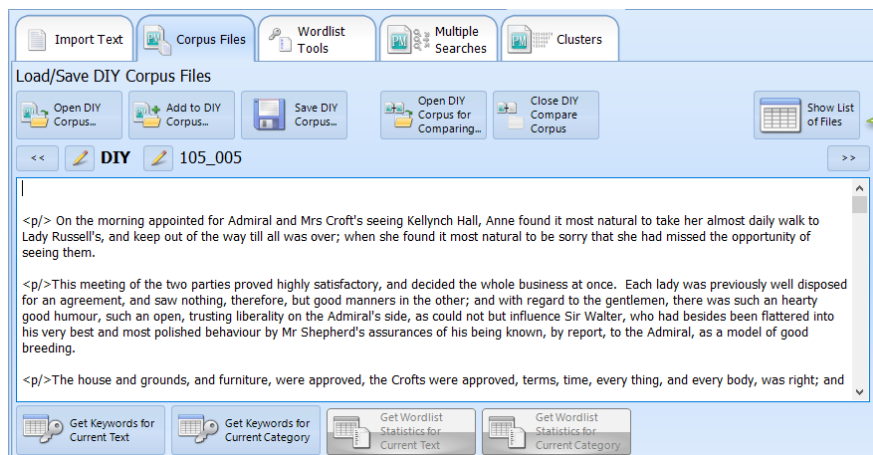
Click the **Show List of Files** button to see filenames and word counts for your DIY corpus.
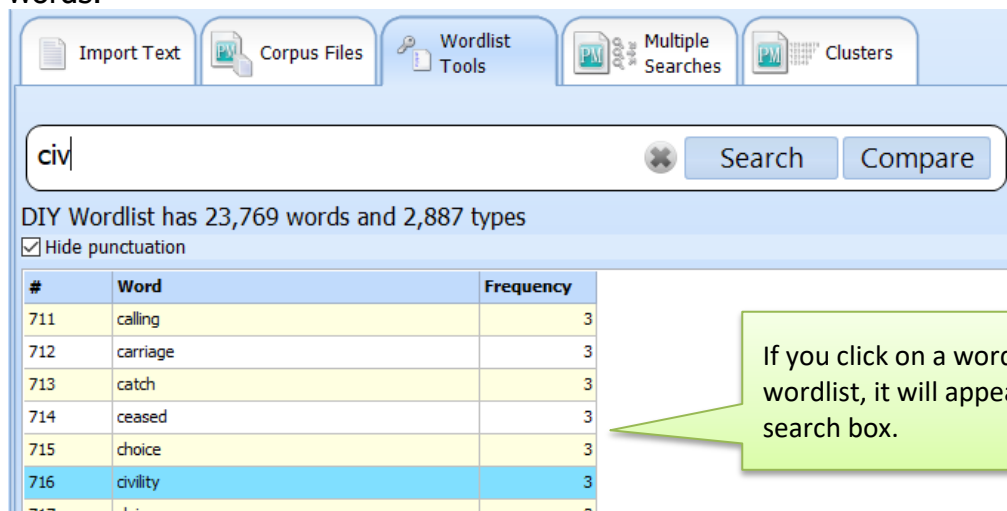

As with other tables in tPM, simply right-click or double-click to copy or save the table as a picture or spreadsheet.

No.	Filename	Category	Token Count
1	105_001	DIY	3,127
2	105_002	DIY	2,298
3	105_003	DIY	3,512
4	105_004	DIY	2,144
5	105_005	DIY	4,009
	Total		15,090

Concordance Cards and Lines (Simple)

For simple concordance searches, you can use the box or wordlist on the Wordlist Tools page. As you type a word into the box, the wordlist will scroll to any matching words.

DIY Wordlist has 23,769 words and 2,887 types

☑ Hide punctuation

#	Word	Frequency
711	calling	3
712	carriage	3
713	catch	3
714	ceased	3
715	choice	3
716	civility	3
717	claims	3

If you click on a word in the wordlist, it will appear in the search box.

As no priming summaries are available for DIY Corpora, multiword searches using ..
are not permitted.

Pattern	Example	Multiple Searches equivalent
Words in a row	kellynch hall	in order with no gaps
Use of _ symbol	at _ kellynch _ hall	in order with no gaps
Use of * symbol	quit * hall	in order (gaps allowed)
Use of | symbol	hall | quit	any order (gaps allowed)

Concordance Cards and Lines (Multiple Searches)

If you want to merge results of two or more words (e.g. *he* or *she*) you can use the
Multiple Searches page. As with the Multiple Searches tool for online corpora, you
can enter one or more words separated in space in up to 5 slots. If you keep the
Ordered List box ticked, this will mean that the words from top to bottom will appear
from left to right in the results.

For example, to merge the results of *he said* with *she said*, you would enter *he she*
in the top box and *said* in the second box. Ticking the No gaps option will ensure no
other words occur in between.

These settings determine the order of the words and whether or not gaps are permitted.

You can change which set of words appears as the **node** by moving this marker. The **node** words always appear in the centre of the concordance lines.

Click these buttons to perform the search.

- When you perform concordance line queries, tPM will calculate scores for Collocations and Links Across Texts so these sorting options will be ready and available on the Lines/Cards tab.
- The only two concordance sorting methods not available for DIY Corpora are Dictionary Style and Links Within Texts.

Collocations

Collocations can be calculated and viewed using the buttons on the Wordlist Tools page. Simply select one of the words on the wordlist and click **Calculate MI Col.** or **Calculate LL Col.**
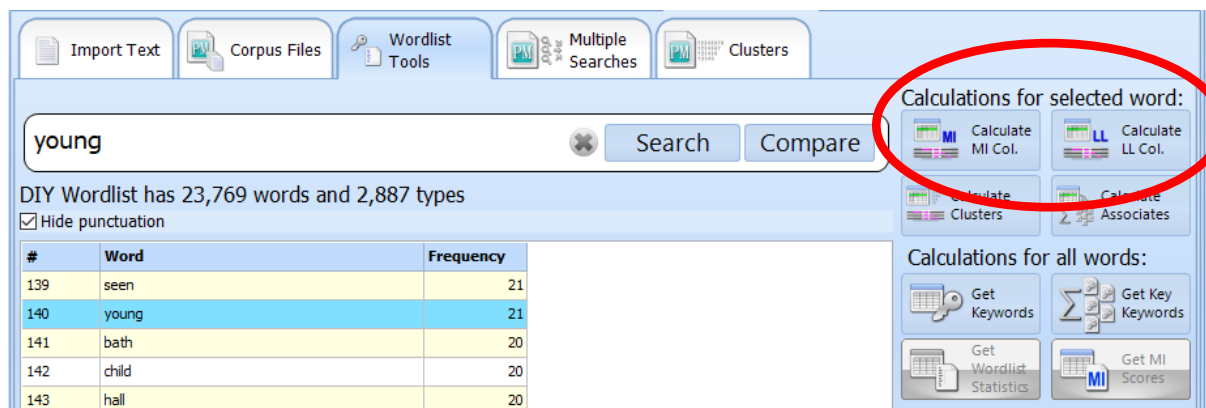
Results for MI Collocations include the collocate, frequency and proportion of instances of each collocate in the 4 word window to the left and right of the node (L4 – R4). You can sort the results by any of the statistical measures by clicking the column headings.

MI Collocations for young (Austen New 3)

Collocate	Freq	L4	L3	L2	L1	R1	R2	R3	R4	T Score	MI3 Score	Dice Score	Delta P	Delta P LR	Delta P RL
the	8	13%	-	13%	63%	-	-	13%	-	2.67	9.48	76.38	0.35	0.30	0.01
a	7	-	29%	43%	14%	-	-	-	14%	2.56	10.01	35.14	0.32	0.27	0.03
man	6	-	-	-	-	100%	-	-	-	2.44	12.99	2.29	0.28	0.00	0.28
and	6	17%	17%	17%	-	17%	17%	-	17%	2.27	8.44	66.57	0.26	0.11	0.11
who	4	-	-	-	-	-	25%	75%	-	1.98	10.39	4.76	0.19	0.00	0.19
so	4	-	-	-	25%	-	25%	-	50%	1.96	9.47	9.33	0.19	0.04	0.14
in	4	-	-	25%	-	-	25%	25%	25%	1.88	7.74	31.81	0.18	0.03	0.13
to	4	-	25%	25%	-	-	-	25%	25%	1.77	6.80	61.52	0.16	0.07	0.07
of	4	-	-	50%	-	-	25%	-	25%	1.76	6.74	64.00	0.16	0.07	0.07
ladies	3	-	-	-	-	100%	-	-	-	1.73	12.90	0.10	0.14	0.00	0.14

Collocations are calculated only within sentence boundaries for tokens up to 4 words to the left or right of the node.

Delta P scores are provided counting instances to the left or right (Delta P), for instances where the collocate is on the left and the node is on the right (Delta P LR) and for instances where the node is on the left and the collocate is on the right (Delta P RL).

As with other tables in tPM, you can copy or save the table as a picture or spreadsheet by right-clicking or double-clicking on it.

LL Collocations are calculated using the same parameters and formula as is used for online corpora in *The Prime Machine*. This Log-likelihood measure uses a different way of calculating collocations from other concordancers, and is based on a measure which takes into account word order and which word is the node.

For DIY Corpora, LL Collocations contain 4 patterns for 2 word collocations only:
- node + collocate (in order with no gap)
- node + gap + collocate (in order with one, two or three words in between)
- collocate + node (in order with no gap)
- collocate + gap + node (in order with one, two or three words in between).

As well as the Log-likelihood score, a Bayes interpretation based on a BIC score is provided to indicate the strength of the collocation. This is sensitive to corpus size.

Clusters (containing a specific word)

Clusters containing a specific word from the wordlist can be counted up and viewed using the button on the Wordlist Tools page. Simply select one of the words on the wordlist and click **Calculate Clusters.**

Results will include clusters of 2, 3, 4 and 5 words (within sentence boundaries), for all clusters with a frequency of 5 or more.
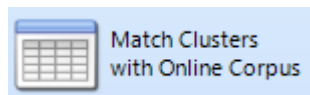
> **Please note:**
> The columns for Categories and Texts in Version 3.0.11.x were not yet operational; Please upgrade to a newer version to calculate these values.

Clusters (across your DIY Corpus)

Rather than viewing clusters containing a predetermined word, you can also generate clusters from across the DIY corpus using the Clusters page. Use the tick boxes to set the length of clusters (from 3 words to 8 words). The selection under the Settings heading will control whether clusters are counted within or across categories. You can also adjust the minimum frequency for clusters here.

Simply click the **Get Clusters** button and wait while the clusters are processed and displayed.

From Version 3.0.14.x, you can also compare frequencies of the clusters from your DIY Corpus clusters with matches in the currently selected Online corpus.

Wordlist Statistics (Online mode only)

When connected to tPM Server you can get total counts of words in your DIY Corpus for the complete list of wordlists available on the Research Tools Wordlists tab.

> Note:
> Some of the wordlists are only available if you care connected to tPM Home Network.

As well as the results for your DIY Corpus, you will be able to see the Log-likelihood and Bayes statistics comparing the counts of words for each wordlist against those for the currently selected corpus. You can select any of the online corpora as the basis for comparison. Simply click the Get Wordlist Statistics button on the Wordlist Tools page.

Wordlist statistics for Austen New 3 compared against BNC Academic
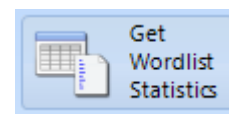
	Wordlist	Study Freq.	Study Per Million	Reference Freq.	Reference Per Million	LL ▽	Bayes
1	Personal Pronouns	2,042	85,910.22	481,408	26,615.90	1887.00	Very strong evidence
2	1st & 2nd Pers. Pronouns	482	20,278.51	101,568	5,615.45	535.70	Very strong evidence
3	Function Words	11,450	481,719.89	7,627,311	421,696.00	134.67	Very strong evidence
4	General Service List 1	16,792	706,466.41	11,449,894	633,038.00	117.72	Very strong evidence
5	Positive words	479	20,152.30	244,749	13,531.60	65.80	Very strong evidence
6	Modals	421	17,712.15	221,854	12,265.80	49.78	Very strong evidence
7	Modals Subgroup 3	113	4,754.09	42,901	2,371.89	43.76	Very strong
8	Modals Subgroup 1	122	5,132.74	69,024	3,816.17	9.69	
9	Modals Subgroup 2	181	7,614.96	109,646	6,062.07	8.68	
10	Negative words	336	14,136.06	336,901	18,626.50	0.00	
11	Archaic Pronouns	0	0.00	426	23.55	0.00	
12	Punctuation	1,119	47,078.13	1,060,127	58,611.90	0.00	
13	General Service List 2	904	38,032.73	750,471	41,49\.80	0.00	
14	Academi... rd List	380	15,987.21	1,367,114	75,584	0.00	

> Results will be sorted by the Log-likelihood score in descending order. These compare your DIY Corpus with the currently selected online corpus.

> Information about the Wordlists can be found on the main Research Tools **Wordlists** page.

> Raw frequencies and frequencies per thousand or per million words are displayed for your DIY Corpus under the columns "Study …"

> Figures for the Reference corpus will be different according to the currently selected online corpus.

If you want to get Wordlist Statistics for the current text or current category, you can use the buttons below the full text display on the Corpus Files page.

[Get Wordlist Statistics for Current Text]   [Get Wordlist Statistics for Current Category]

> 💡 • You can switch between seeing relative frequencies per thousand or per million using a setting on the Tables page of the Options Tab.

Average Mutual Information Scores (Online mode only)

When connected to tPM Server you can get average scores of the collocations for each text in your DIY Corpus. This function does take some time to complete as each pair of words in a five word window (bound by sentence boundaries) has to be checked against the currently selected online corpus for mutual information collocation scores.

Simply click the **Get MI Scores** button on the Wordlist Tools page.
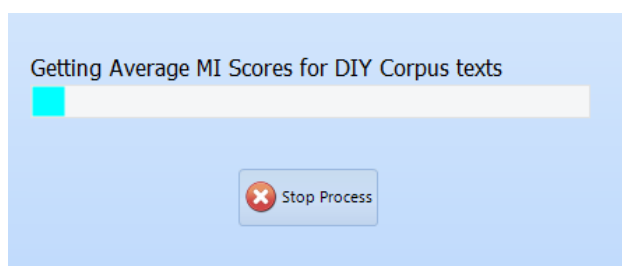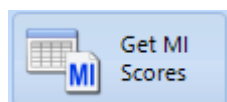
Pairs of words will be sent to the server in batches.

If you stop the process, partial results will be displayed (the word "incomplete" will appear in the heading).

Results will be displayed for each text and for the whole DIY corpus.

Average MWU scores for Austen New 3 compared against Fiction Collection 37x1

	Filename	T Score	MI3 Score	Dice Score	Delta P	Delta P ↔
7	105_007	41.44	19.68	0.04	0.11	0.04
6	105_006	33.10	19.18	0.03	0.10	0.03
5	105_005	41.07	19.98	0.04	0.12	0.03
4	105_004	31.93	18.90	0.03	0.13	0.04
3	105_003	31.74	16.95	0.03	0.07	0.03
2	105_002	40.93	18.34	0.04	0.11	0.03
1	105_001	34.32	18.51	0.03	0.12	0.03
	Austen New 3	**35.52**	**18.62**	**0.03**	**0.10**	**0.03**

Scores for each statistic are based on the scores for each pair of words in the currently selected online corpus. Pairs of "words" will also include punctuation. Therefore, if you choose a different reference corpus the results will be different.

For example:

    Sentence: *The cat sat on the mat.*
    Pairs: the + cat; the + sat; the + on; the + the
             cat + sat; cat + on; cat + the; cat + mat
             sat + on; sat + the; sat + mat; sat + .
             on + the; on + mat; on + .
             the + mat; the + .
             mat + .

The T Scores, MI3 Scores and Dice Scores are symmetrical for each node and collocate, so the average scores for these are simply calculated by taking the average for each combination of word pairs.

The Delta P Scores are asymmetrical, so for the purposes of this process the higher of the two Delta P scores is taken. This means in the example above, we would expect that the Delta P score for *the + cat* would use *cat* as the node and *the* as the collocate (as we can expect this to be higher than the Delta P score for *the* as a node).

As well as a Delta P Score for the words in each pair in any order, the Delta P ↔ column gives scores where the order of words within the 5 word window is taken into account.

Using a Second DIY Corpus as a Reference Corpus

The remainder of this chapter provides information about DIY Corpus Tools which can be used to compare your DIY Corpus with an online reference corpus **or** to compare one DIY Corpus with another previously saved DIY Corpus.

In order to use the DIY Corpus tools listed below, you will need to either go online or open a previously saved DIY Corpus for comparing. If you are online, these tools will always use the online corpus as a reference. Be sure to set tPM to work offline if you want to use your second DIY corpus as a reference corpus.
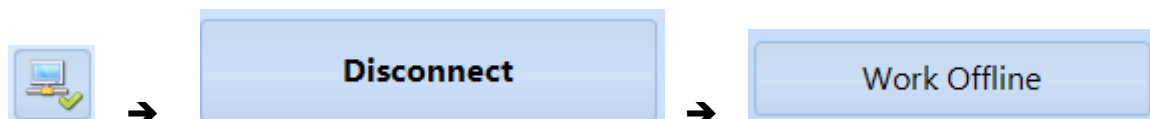
The buttons to Open or Close a DIY Compare Corpus are on the **Corpus Files** page.
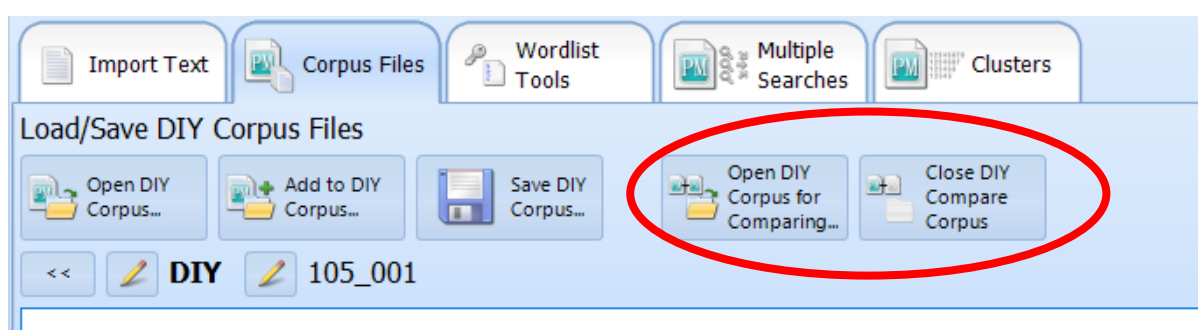
When you open a DIY Corpus for comparing, you cannot edit its contents. If you are starting two new DIY corpora and you want to compare them you should:

1. Create the DIY Corpus you want to use as a reference corpus first. Import the files and save it.
2. Create the DIY Corpus you want to use as your study corpus next. Import the files and save it.

3. Use the **Open DIY Corpus for Comparing...** button to reopen the corpus you saved in Step 1.
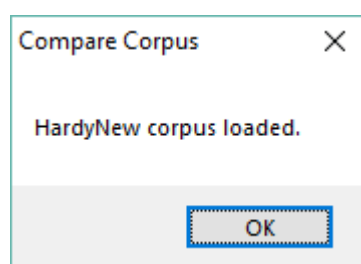
When you open a DIY Corpus for Comparing, you will get dialog box confirmation, but you won't be able to see any specific information about the corpus.

Comparing Concordance Lines from Two Corpora

When working online or if you have opened a DIY Corpus for comparing, you can view concordance lines for your main DIY corpus compared to a reference corpus using the **Compare** buttons on the **Wordlist Tools** and **Multiple Searches** pages.
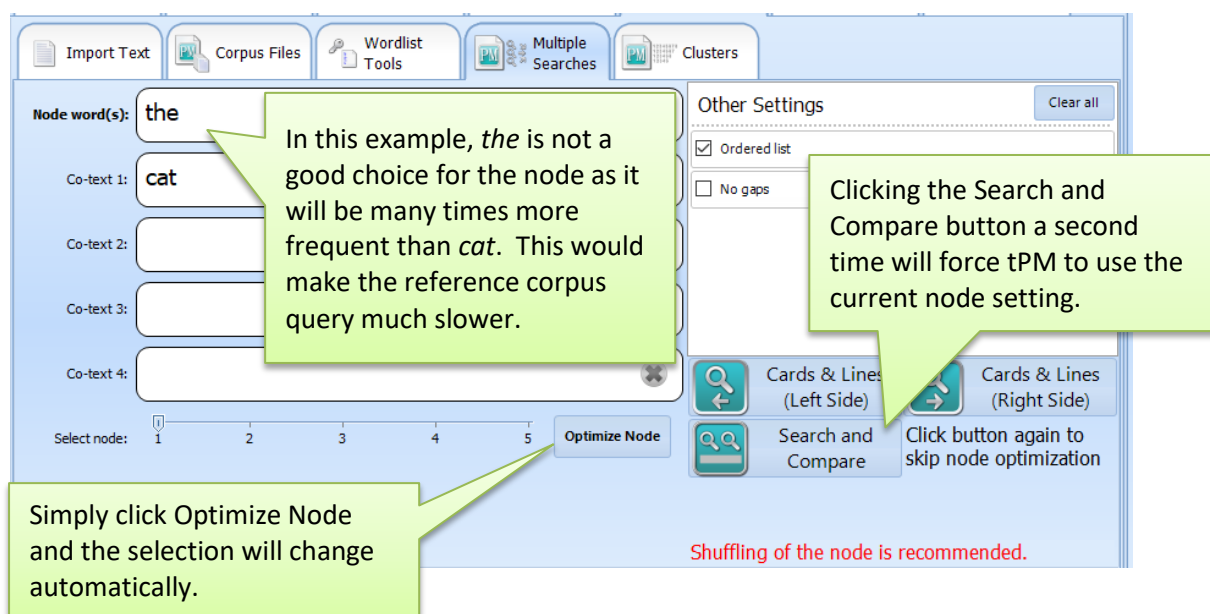
These buttons allow you to see results for the same queries with your main DIY Corpus on the left and the reference corpus on the right.

When using tPM online corpora for comparison, the node word will be determined for the online corpus and matched in the DIY Corpus.

From the Wordlist Tools page, this means that if you enter two or more words, the online reference corpus will be checked and the word with the lowest frequency will be automatically selected as the node.

From the **Multiple Searches** page for DIY Text Tools, if you enter words in two or more of the boxes and the box labelled **Node word(s):** does not contain words with the lowest frequency, you will be prompted and encouraged to Optimize the Node for this query.

In this example, *the* is not a good choice for the node as it will be many times more frequent than *cat*. This would make the reference corpus query much slower.

Clicking the Search and Compare button a second time will force tPM to use the current node setting.

Simply click Optimize Node and the selection will change automatically.

Clicking the Optimize Node button will simply change the node selection based on frequency information from the online corpus. You can override this by clicking the **Search and Compare** button a second time, but the response time for the reference corpus query may be much slower.
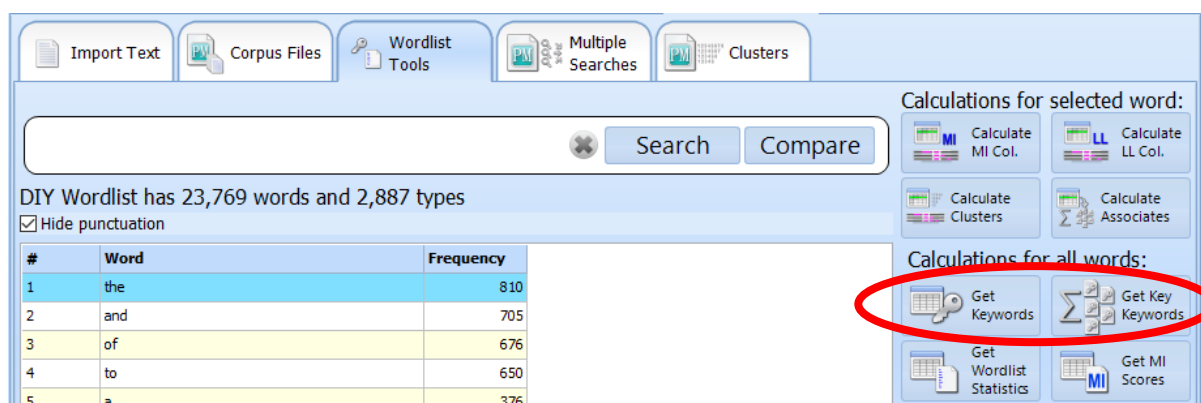
Key Words and Key Key Words

When working online or if you have opened a DIY Corpus for comparing, you can easily calculate and view Key Words and Key Key Words for your DIY Corpus.

The Key Words buttons give results for frequencies of words across your whole DIY Corpus. Key Key Words take each text in your corpus one by one and generate Key Words for these, before combining the results to show words which are Key Words in 2 or more texts.

To use the Key Key Words buttons you will need, of course, to have 2 or more texts in your DIY Corpus.
Key Words and Key Key Words for the whole DIY Corpus can be calculated using the buttons on the Wordlist Tools page.

You can also calculate Key Words for a single text or for frequencies of words across one category using the buttons below the full text display on the **Corpus Files** page.
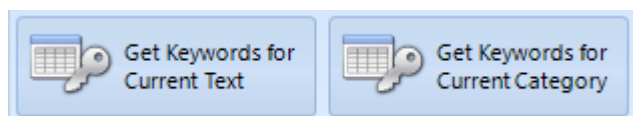
Key Words and Key Key Words are based on the log-likelihood scores comparing the frequencies of words in your DIY Corpus and the frequencies of these words in the reference corpus, taking into account the overall sizes of these corpora. Therefore, if you use a different corpus as a reference corpus, you will get different results.

Key Associates

Key Associates are calculated by finding which other words are Key Words in the texts in which a specific word is a Key Word. Key Associates show you some of the other words which are often used in the texts where the word you have selected is repeated.

To calculate Key Associates for a specific word in your DIY Corpus, select the word on the Wordlist Tools page and then click the **Calculate Associates** button.
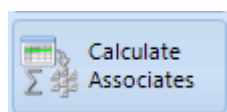
Key Associates are based on Key Word scores for each text in your DIY Corpus compared against the reference corpus (the currently selected tPM online corpus or a second DIY Corpus opened for comparing).

Support

The Prime Machine is still undergoing development.
For further information see http://help.theprimemachine.com

Last Updated: 26 October 2018