

Keywords Analysis notebook – help pages

Introduction

The Keyword Analysis tool is a Jupyter notebook containing code that was developed by the [Sydney Informatics Hub](#) (SIH) in collaboration with the [Sydney Corpus Lab](#) as part of the [Australian Text Analytics Platform](#) (ATAP) project. The tool is designed to analyse words in two (or more) corpora and identify whether certain words are over- or under-represented in the ‘node’ or ‘study’ corpus (i.e., the corpus of interest) compared to a ‘reference’ corpus (i.e., the standard of comparison). It uses the keywords analysis technique pioneered by Mike Scott (Scott, 1997) and draws on statistical formulae provided by Paul Rayson (see below).

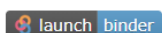
(Note: if you are unfamiliar with how to use Jupyter Notebooks, have a look at this [guide](#).)

Getting started

The tool is available on [GitHub](#) where you can launch the tool on Jupyter Notebook via Binder by clicking on one of the ‘launch binder’ buttons:

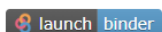
Setup

This tool has been designed for use with minimal setup from users. You are able to run it in the cloud and any dependencies with other packages will be installed for you automatically. In order to launch and use the tool, you just need to click the below icon.



Note: CILogon authentication is required. You can use your institutional, Google or Microsoft account to login.

If you do not have access to any of the above accounts, you can use the below link to access the tool (this is a free Binder version, limited to 2GB memory only).



It may take a few minutes for Binder to launch the notebook and install the dependencies for the tool. Please be patient.

The access to the ATAP Binderhub (i.e., the first ‘launch binder’ button) requires CILogon authentication, which supports the single sign-on (SSO) method with most (Australian or international) institutional login credentials, ORCID, or a Google/Microsoft account. If you have trouble authenticating, please refer to the [CILogon troubleshooting guide](#). If you have access to software that supports Jupyter Notebooks, you can also clone the Github repository and use the notebook locally (i.e., without Internet connection) on your own computer.

Overview of Tool

If you have already read [the blog post introducing this tool](#) or are familiar with the tool, you can skip this general overview section and will find the tool user instructions from [here](#) onwards.

The tool allows you to upload texts as individual files or as corpora/datasets (e.g., in zipped files). You can also upload frequency lists instead of the full text corpora if those are already generated by another software. Once your texts have been uploaded, several statistical calculations will be applied on words in your corpus/corpora; these include python implementations of methods provided on Paul Rayson's [website](#) (e.g. Log Likelihood, %Diff, Bayes Factor, Effect Size for Log Likelihood, Relative Risk, Log Ratio and Odds Ratio). This notebook re-engineering is conducted with permission; for statistical formulae and explanations with relevant attribution please refer to the link above.

You can conduct keyword statistics between pairs of datasets (i.e., node corpus vs. reference corpus). If you have uploaded more than two corpora (e.g., corpus 1, corpus 2 and corpus 3), you can choose to compare one corpus to another (e.g., using corpus 1 as node corpus and comparing it to corpus 2 as the reference corpus), or a selected corpus against all other corpora as reference (e.g., using corpus 2 as the node corpus and comparing it to the reference corpus which is comprised of corpus 1 and 3). You can then visualise the statistics from the keyword analysis as a line graph (an example is shown in Figure 1 below).

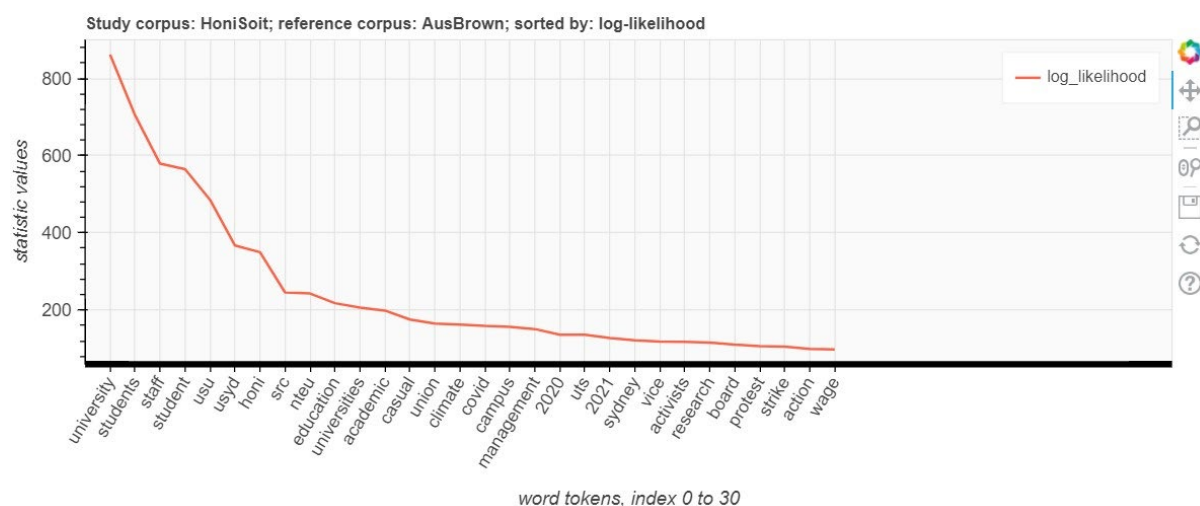


Figure 1. Top 30 positive keywords (x-axis) with the highest log-likelihood values (y-axis)

Using default settings, the resulting line graph (e.g., Figure 1) will show 30 positive keywords (i.e., those that are 'overused' in the node/study corpus in comparison to the reference corpus) with the highest log-likelihood values. You can modify the resulting visualisation by:

- Selecting one or more statistics to display (log-likelihood, percent-difference, Bayesian information criterion, etc.);
- Choosing how to sort the keywords (alphabetically or according to a selected statistical value);
- Choosing which type of keywords to display (i.e., positive, negative, or all keywords) and adjusting the number of displayed keywords.

You can save and download the visualisation to your computer using the 'save' icon on the right-hand side of the graph (see right side of Figure 1). You can also save the results of the statistical analysis of keywords into an excel spreadsheet and download it to your local computer.

Additionally, the tool allows you to compare words across multiple corpora. For this multi-corpus analysis, the statistics are calculated across the whole set in one go: An adjusted/expected average is calculated for each corpus and actual/observed frequencies are then compared to this average. However, please note that only some of the statistics (i.e., log-likelihood, Bayes factor BIC, and effect size for log-likelihood) can be used to conduct the multi-corpus analysis, since the other statistics can only be used for pairwise comparisons.

The notebook also goes beyond keyword analysis in allowing you to investigate if the use of a particular word in a corpus is statistically different to the use of that same word in a different corpus. All you need to do is enter the word (e.g., *university* in Figure 2 below) you wish to analyse, select the two corpora you wish to compare, perform data transformation if needed (i.e., using log transformation or square root transformation), and select the statistical test you want to use (i.e., Welch t-test or Fisher permutation test). Please note that you will not be able to conduct these statistical tests if you have uploaded frequency word lists (instead of individual files or as corpora/datasets).

The screenshot shows a web-based interface for statistical analysis. It includes three main input sections: 'Enter the word you wish to analyse' with a text box containing 'university'; 'Data transformation:' with a dropdown menu set to 'no transform'; and 'Select statistical test:' with a dropdown menu set to 'Fisher Permutation test'. Below these are two buttons: 'Plot histogram' and 'Perform statistical analysis'. A second section, 'Enter the name of the two corpora you wish to compare', contains two text boxes: 'Corpus 1:' with 'HoniSoit' and 'Corpus 2:' with 'AusBrown'. At the bottom, a pink-shaded output area displays the following text: '100% ██████████ | 269/269 [00:00<00:00, 16534.55it/s] Fisher Permutation test', 'Statistic score: 6.81', 'p-value: 0.00', 'The mean frequency of the word 'university' is higher in 'HoniSoit' than in 'AusBrown', and we consider the difference to be statistically significant.', and 'In summary, we reject the null hypothesis that use of the word 'university' in 'HoniSoit' is equal to that in 'AusBrown'.'

Figure 2. Statistical analysis of *university* using default settings

Setup

Before you begin, the KeywordAnalysis package and the necessary libraries need to be imported and initiated; these are all pre-configured in the first executable cell of the notebook.

1. Execute the cell (as a reminder, you can do this by pressing 'Ctrl' + 'Enter' on PC or 'Cmd' + 'Enter' on Mac):


```
[ ]: # import the KeywordsAnalysis tool
print('Loading KeywordsAnalysis...')
from keywords_analysis import KeywordsAnalysis, DownloadFileLink

# initialize the KeywordsAnalysis
ka = KeywordsAnalysis()
print('Finished loading.')
```

2. Once completed, you should see a printed message, "Finished loading", as shown below:

```
[1]: # import the KeywordsAnalysis tool
print('Loading KeywordsAnalysis...')
from keywords_analysis import KeywordsAnalysis, DownloadFileLink

# initialize the KeywordsAnalysis
ka = KeywordsAnalysis()
print('Finished loading.')
```

Loading KeywordsAnalysis...
 BokehJS 2.4.3 successfully loaded.

[nltk_data] Downloading package punkt to /home/jovyan/nltk_data...
 [nltk_data] Unzipping tokenizers/punkt.zip.
 Finished loading.

Load the data

This notebook will allow you to upload corpus text data in one or multiple text files or an excel spreadsheet.

If you wish to upload your corpora all at once in an Excel spreadsheet, please format the header of your spreadsheet as in the following example:

text_name	text	source
text1	Facebook and Instagram, which Facebook owns, followed up in the evening, announcing that Trump wouldn't be able to post for 24 hours following two violations of its policies. The White House did not immediately offer a response to the actions. While some cheered the platforms' response, experts noted that these actions follow years of hemming and hawing regarding Trump and his supporters spreading dangerous misinformation and encouraging violence that contributed to Wednesday's events.	corpus1
text2	(CBC News) Republican lawmakers and previous administration officials had begged Trump to give a statement to his supporters to quell the violence. He posted his video as authorities struggled to take control of a chaotic situation at the Capitol that led to the evacuation of lawmakers and the death of at least one person. Lawmakers, world leaders condemn chaos at the U.S. Capitol while some call for Trump's removal Trudeau says Canadians 'deeply disturbed' by violence in Washington D.C.	corpus1

Alternatively, you can upload a frequency word list. You can store the word frequencies in an Excel spreadsheet. The spreadsheet should follow a format similar to the one shown below (i.e., the first column should contain the words and the second column the word frequencies):

word	freq
I	11653
YOU	11617
THE	7388
S	5922
TO	5809


1. Execute the cell:

```
[ ]: # upload the text files and/or excel spreadsheets onto the system  
ka.upload_file_widget()
```

2. Since you need to upload at least two corpora, give a name to each corpus first by typing it into the box next to “Corpus Name”:

```
[2]: # upload the text files and/or excel spreadsheets onto the system  
ka.upload_file_widget()
```

[2]: Corpus Name: Uploading word frequency list

 Upload your files (txt, csv, xlsx or zip) (0)

Uploading large files may take a while. Please be patient.
Please wait and do not press any buttons until the progress bar appears...

If you are uploading a frequency list, make sure you check the “Uploading word frequency list” box.

3. Click ‘Upload your files’. A window should appear prompting you to select txt files, or a single csv file, xlsx file, or zip folder.
4. Click ‘Open’ after you’ve selected the file(s) you want to upload.
5. The tool should start loading the selected file(s). **Be patient and wait for the process to finish even if it looks like nothing is happening – it can take a while.** Once completed, you should get a message saying, “Finished uploading files...” and another message describing the number of files that were uploaded. For example:


```
[2]: # upload the text files and/or excel spreadsheets onto the system
ka.upload_file_widget()
```

[2]: Corpus Name: Uploading word frequency list

 Upload your files (txt, csv, xlsx or zip) (0)

Uploading large files may take a while. Please be patient.
Please wait and do not press any buttons until the progress bar appears...

The total size of the upload is 0.38 MB.
Reading uploaded files...
This may take a while...

100%  | 100/100 [00:00<00:00, 39423.86it/s]

Finished uploading files.
100 text documents are loaded.

You can now upload your next corpus, or continue to the next step

6. You can now upload another corpus. Repeat steps 2-5 for every additional corpus you want to upload.

Calculate word statistics



Once your texts have been uploaded, you can begin to calculate the statistics for all of the words in the corpus. If the corpora are not uploaded as pre-processed frequency list, the tool will also automatically tokenise your corpora using the [CountVectorizer](#) from the Scikit-learn python library.

1. Execute the cell:

```
[ ]: # begin the process of calculating word statistics
ka.calculate_word_statistics()
```

2. Once completed, you should get progress bars displaying 100%, similar to the ones shown below:

```
[3]: # begin the process of calculating word statistics
ka.calculate_word_statistics()
```

Step 1/3: 100%  | 2/2 [00:00<00:00, 5.83it/s]
Step 3/3: 100%  | 2/2 [00:00<00:00, 2.23it/s]

Analyse word statistics

Once the tool has finished calculating the statistics, you can analyse and visualise the outcome.

Pairwise analysis

You can use the tool to analyse keyword statistics between pairs of datasets (node/study corpus vs. reference corpus). When you have more than two datasets to compare (e.g., corpus 1, corpus 2, and corpus 3), you can either choose to compare one corpus to another

(e.g., corpus 1 as the node/study vs. corpus 2 as the reference corpus), or one corpus with the rest of the data (e.g., study corpus: corpus 2 vs reference corpus: rest of corpus, which includes corpus 1 and 3). The tool will produce a line graph to display the results of the pairwise keyword analysis.

1. Execute the following cell:

```
[ ]: # generate pair-wise corpus analysis
ka.analyse_stats(right_padding=0.9) # adjust the 'right_padding' to move the Legend box left/right
```

2. Once completed, you should get several widgets to adjust the settings for the line graph:

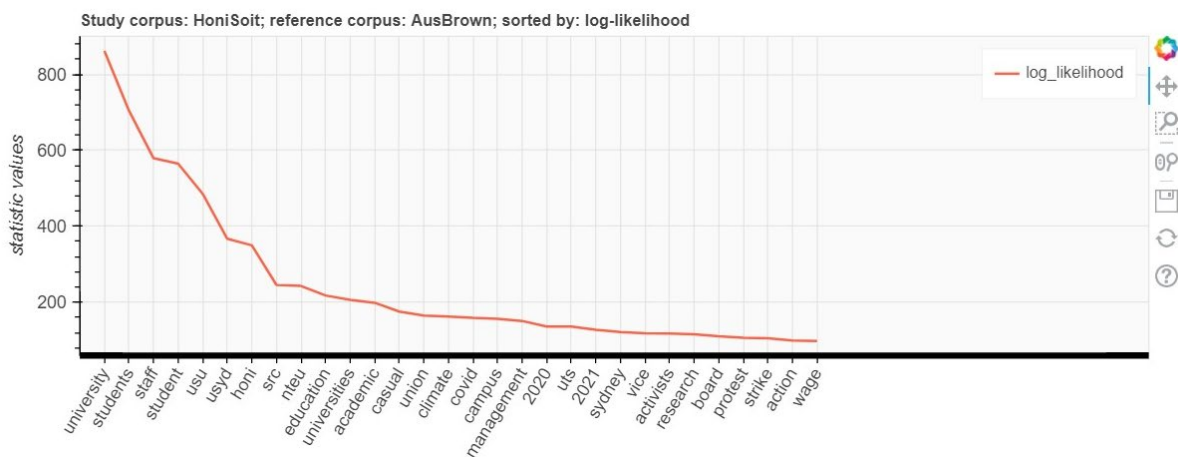
```
[4]: # generate pair-wise corpus analysis
ka.analyse_stats(right_padding=0.9) # adjust the 'right_padding' to move the Legend box left/right
```

[4]:

<p>Select study corpus:</p> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">HoniSoit ▼</div> <p>Select reference corpus:</p> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">AusBrown ▼</div>	<p>Select statistic(s) to display:</p> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;"> normalised word count (study corpus) ▲ normalised word count (reference corpus) ▲ log-likelihood percent-diff ▼ </div>	<p>Sorted by:</p> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">log-likelihood ▼</div> <p>Keywords to display:</p> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">positive (overuse) ▼</div>	<p>Select index:</p> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px; width: 50px;">0</div> <div style="text-align: center; margin-top: 10px;"> <div style="border: 1px solid #ccc; padding: 5px; width: 100px; margin: 0 auto; background-color: #f0f0f0;">Display chart</div> <div style="border: 1px solid #ccc; padding: 5px; width: 100px; margin: 5px auto; background-color: #f0f0f0;">Save data to excel</div> </div>
---	--	---	---

3. Select the node/corpus by selecting one of the corpora you’ve uploaded from the “Select study corpus” drop-down menu. As a reminder, this will be the corpus for which you are retrieving keywords.
4. Select the reference corpus by selecting one of the other corpora you’ve uploaded from the “Select reference corpus” drop-down menu. As a reminder, this corpus is the one you’re comparing the node/study corpus to (i.e., the keywords retrieved aren’t for this corpus). You also have a “rest of corpus” option which effectively combines corpora that you’ve uploaded (excluding the one you selected as the node/study corpus) into one reference corpus.
5. Select the statistics to be displayed in the y-axis of the resulting line graph. The default option is “log-likelihood” which means the resulting graph will display the statistical significance (i.e., log-likelihood) value for the frequency difference across the two corpora. Other options include the normalised word count (i.e., normalised frequency) in the node corpus, normalised word count in the reference corpus, percent-diff (i.e., percentage difference), bayes factor BIC (i.e., the Bayesian information criterion), effect size for log-likelihood (ELL), relative risk, log ratio, and odds ratio (see [here](#) for more information about these statistics). You can select multiple statistics by using the left-click button on your mouse while holding the Ctrl or Command button.
6. Select how the keywords will be displayed. This option effectively sorts the keywords based on the selected statistic value (e.g., normalised word count, log-likelihood, percent-diff) from highest to lowest or in alphabetical order. The default option is “log-likelihood” which means the tool will display the keywords with the highest log-likelihood in the resulting line graph.

7. Select which type of keywords to display. The default option is “positive (overuse)” which means that the tool will only display positive keywords (i.e., those that occur more frequently in the node corpus than in the reference corpus). You can also choose to display negative keywords (i.e., those that occur less frequently in the node corpus than in the reference corpus) or all keywords (i.e., both positive and negative keywords).
8. By default, the tool will only display the keywords in the index 0 to 30 positions (i.e., keywords ranked 1 to 31) in the x-axis of the resulting graph. You can view a different set of words by using the up/down arrow in the box under “Select index”. For example, if you click up, the index will shift from 0-30 to 10-40, which means keywords ranked 11-41 will be displayed instead. You can also enter a number that’s not a multiple of 10 (e.g., 1, 2, 3, ...) and click 'Tab' (on your keyboard).
9. Once you’re happy with the settings, click “Display chart” and you should get a line chart showing the keywords for the selected study/node corpus. The example graph below uses most of the default settings (i.e., Statistics: log-likelihood, Sorted by: log-likelihood, Keywords to display: Positive, Select index: 0):



You can also zoom in/out of the line graph, download/save the visualisation to your local computer using the icons in the interactive toolbar on the right hand-side of the graph.

10. Click on “Save data to Excel” to download and save the results of the pairwise comparison. You will get a table previewing the content of the Excel file (containing e.g., words occurring in either the node or reference corpus, frequency of the word in the node corpus, frequency in the reference corpus, etc.) and a link to the generated Excel spreadsheet file:

Please generate and display a chart first before saving!

	HoniSoit	AusBrown	word	total_word_used	expected_study_corpus_wc	expected_reference_corpus_wc	normalised_study_corpus_wc	normalised_reference_corpus_wc	overuse_word_in_study_corpus	log
0	5	9	0	14	3.320163	10.679837	0.000081	0.000045	True	
1	0	6	00	6	1.422927	4.577073	0.000000	0.000030	False	
2	49	202	000	251	59.525783	191.474217	0.000795	0.001019	False	
8	2	0	00pm	2	0.474309	1.525691	0.000032	0.000000	True	
26	0	2	019	2	0.474309	1.525691	0.000000	0.000010	False	

Saving in progress...
 Saving is complete.
 Click below to download:
 study_HoniSoit_ref_AusBrown.xlsx

11. Click on the xlsx file to start the download.

Multi-corpus analysis

You can also use the tool to analyse the overall statistics at the multi-corpora level, for cases where you explore more than two corpora. For this multi-corpus analysis, the statistics are calculated across the whole set in one go: an adjusted/expected average is calculated for each corpus and actual/observed frequencies are then compared to this average. Only some of the statistical measures are used (details are provided in the notebook itself and also above). The tool will produce a line graph to display the results of the multi-corpora keyword analysis.

1. Execute the following cell:

```
[ ]: # generate multi-corpus analysis
ka.analyse_stats(right_padding=0.5, multi=True)
```

2. Once completed, you should get several widgets to adjust the settings for the line graph:

```
[5]: # generate multi-corpus analysis
ka.analyse_stats(right_padding=0.5, multi=True)
```

[5]: **Select statistic(s) to display):**

log-likelihood
 bayes factor BIC
 ELL

Sorted by:

log-likelihood

Select index:

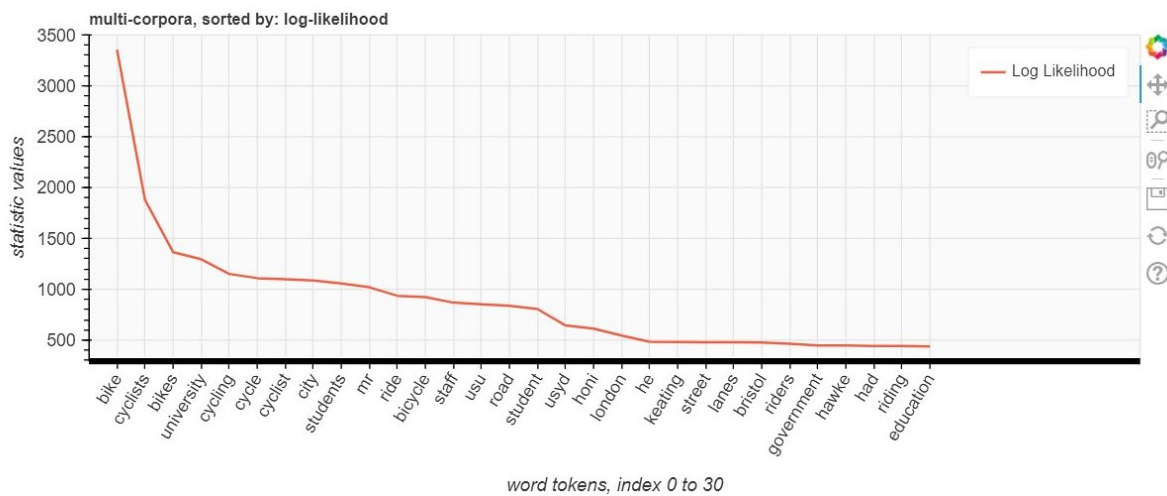
0

Display chart

Save data to excel

3. Select statistics to display. The default option is “log-likelihood”. Other options include bayes factor BIC (i.e., the Bayesian information criterion) and effect size for log-likelihood (ELL). You can select multiple statistics by using the left-click button on your mouse while hold the Ctrl or Command button.
4. Select how the keywords will be displayed. This option effectively sorts the keywords based on the selected statistic value (i.e., log-likelihood, bayes factor BIC, and ELL) from highest to lowest or in alphabetical order. The default option is “log-likelihood” which means the tool will display the keywords with the highest log-likelihood in the resulting line graph.

- Like with the pairwise comparison, you can select the index by using the up/down arrow in the box under “Select index”. By default, the tool will only display the keywords in the index 0 to 30 positions (i.e., keywords ranked 1 to 31) in the resulting graph. As a reminder, if you click up, the index will shift from 0-30 to 10-40, which means keywords ranked 11-41 will be displayed instead. You can also enter a number that’s not a multiple of 10 (e.g., 1, 2, 3, ...) and click 'Tab' (on your keyboard).
- Once you’re happy with the settings, click “Display chart” and you should get a line chart showing the keywords for the multi-corpus comparison. The example graph below uses the default settings (i.e., Statistics: log-likelihood, Sorted by: log-likelihood, Select index: 0):



You can also zoom in/out of the line graph, download/save the visualisation to your local computer using the icons in the interactive toolbar on the right hand-side of the graph.

- Click on “Save data to Excel” to download and save the results of the comparison. You will get a table previewing the content of the Excel file (containing e.g., words occurring in any of the corpora you’ve uploaded, frequency of the word in each corpus, etc.):

	HoniSoit	AusBrown	Cycling	word	total_word_used	expected_wc_AusBrown	expected_wc_Cycling	expected_wc_HoniSoit	Log Likelihood	Bayes Factor	BIC	ELL
0	5	9	11	0	25	6.386712	16.627778	1.985510	6.319764	-20.803330	1.187819e-05	
1	0	6	0	00	6	1.532811	3.990667	0.476522	16.375875	-10.747218	-2.848003e-05	
2	49	202	712	000	963	246.016135	640.502011	76.481854	27.422583	0.299490	8.150957e-06	
3	0	0	2	000ft	2	0.510937	1.330222	0.158841	1.631205	-25.491889	-1.142932e-06	
4	0	0	1	000km	1	0.255468	0.665111	0.079420	0.815602	-26.307492	-4.150862e-07	

Saving in progress...
 Saving is complete.
 Click below to download:
[multi_corpora_analysis.xlsx](#)

- Click on the xlsx file (called “multi_corpora_analysis.xlsx” by default) to start the download.

Welch t-test and Fisher permutation test

Finally, you can also use the Welch t-test and the Fisher permutation test to investigate if the use of a certain word in a corpus is statistically different to the use of that same word in a different corpus.

Please note that you won't be able to perform these statistical tests if you only upload word frequency lists, as the analysis is conducted on the number of words in each text within the corpus. You will need to upload the actual text files to do this.

1. Execute the cell:

```
[ ]: ka.word_usage_analysis()
```

2. Once completed, you should get several widgets to adjust the settings for the statistical analysis:

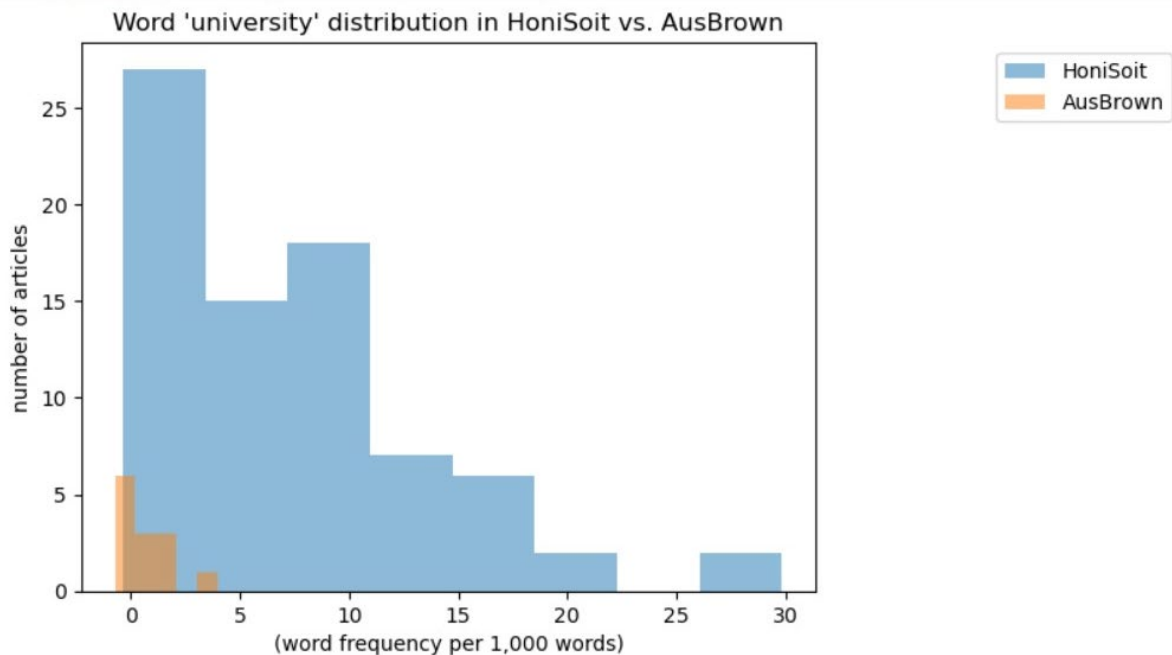
```
[6]: ka.word_usage_analysis()
```

100% | ██████████ | 269/269 [00:03<00:00, 74.22it/s]

[6]: Enter the word you wish to analyse

Word: <input type="text" value="Enter a word..."/>	Data transformation: <input type="text" value="no transform"/>	Select statistical test: <input type="text" value="Fisher Permutation test"/>
Enter the name of the two corpora you wish to compare	<input type="button" value="Plot histogram"/>	<input type="button" value="Perform statistical analysis"/>
Corpus 1: <input type="text" value="Enter name of corpus 1..."/>		
Corpus 2: <input type="text" value="Enter name of corpus 2..."/>		

3. Enter the word you wish to analyse in the box next to “Word”. Only enter single words (e.g., *university*), **not** phrases or multiword expressions (e.g., *the University of Sydney*).
4. Select which two of the corpora among the ones you've uploaded you wish to compare.
5. Select the type of data transformation. The default option is “no transformation” which means your data won't be transformed to resemble [normal distribution](#). You can choose to apply log transformation or square root transformation to your data. You can click on “Plot histogram” to see what your data looks like without any transformations or with one of the transformations:



6. Select whether you want to use the Welch t-test or the Fisher permutation test.
7. Once you're happy with the settings, click "Perform statistical test". You will get the statistical score and the p-value from the statistical test. You will also get a message indicating which corpus contains a higher mean frequency of the word, whether this difference in mean frequency is significant, and whether the null hypothesis (i.e., the use of the word is equal in both corpora) is confirmed or rejected. For example:

```
100%|██████████| 269/269 [00:00<00:00, 25484.33it/s]
Fisher Permutation test
Statistic score: 6.81
p-value: 0.00

The mean frequency of the word 'university' is higher in 'HoniSoit' than in 'AusBrown',
and we consider the difference to be statistically significant.

In summary, we reject the null hypothesis that use of the word 'university' in 'HoniSoit' is equal to that in 'AusBrown'.
```

If you've entered a word that does not occur in either corpus or a multiword expression, you will get the following message instead:

```
100%|██████████| 1956/1956 [00:00<00:00, 43728.63it/s]
The word 'permutation' does not exist in the selected corpora.
```

Citing/Referencing this Notebook

Citation: Jufri, S., & Sun, C. (2022). Keywords Analysis (version 1.0) [Jupyter notebook]. Australian Text Analytics Platform. <https://github.com/Australian-Text-Analytics-Platform/keywords-analysis>

You can adjust the year and version number in the above citation depending on the version of the notebook that you have used in your study.

If you are using this notebook in your research, please also include the following statement or an appropriate variation thereof:

This study has utilised a notebook/notebooks developed for the Australian Text Analytics Platform (<https://www.atap.edu.au>) available at <https://github.com/Australian-Text-Analytics-Platform/keywords-analysis>.

In addition, please inform ATAP (info@atap.edu.au) of publications and grant applications deriving from the use of any ATAP notebooks in order to support continued funding and development of the platform.

Acknowledgments

This Jupyter notebook and relevant python scripts were developed by the Sydney Informatics Hub (SIH) in collaboration with the Sydney Corpus Lab under the [Australian Text Analytics Platform program](#) and the [HASS Research Data Commons and Indigenous Research Capability Program](#). These projects received investment from the Australian Research Data Commons ([ARDC](#)), which is funded by the National Collaborative Research Infrastructure Strategy ([NCRIS](#)).

Known Issues

The notebook has not been tested with very big data sets.