

# Concordancer Tool notebook – help pages

## Introduction

The Concordancer 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 allows users to search a text/corpus for every instance of a search term and then presents the found instances in the form of a concordance. This Concordancer has specifically been designed to allow users to

- undertake ‘dialogic’ analysis (when the input consists of related text pairs, such as question-answer or social media post-response) and/or
- analyse the meta-data that are associated with each occurrence of the search term, if such meta-data are included in the input (for example, speaker identity, political affiliation, company, and so on).

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

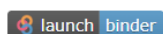
## Getting started

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:

### Standalone tools

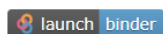
---

Use the Concordancer as a standalone tool by clicking the following Binder link



Note: CILogon authentication is required. You can use your institutional, Google or Microsoft account to login. If you have trouble authenticating, please refer to the [CILogon troubleshooting guide](#).

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).



The access to the ATAP Binderhub (i.e., the first ‘launch binder’ button) requires CILogon authentication, which supports single sign-on (SSO) method with most (Australian or international) institutional login credentials or a Google/Microsoft account. 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

Like existing concordancers, the ATAP Concordancer tool allows you to upload text data (e.g., as a .csv or .txt file), search the text for each instance of a search term, and presents the instances in the form of a concordance:

	left_context	match	right_context	text_id
0	what is the biggest change that you've noticed in	corpus	research throughout your career?	0
1	ch was amazing. 100-million-word British National	Corpus	– wow! Now we have billions of words that we can	1
2	s preparing for that, I wanted to look at applied	corpus	linguistics and looked through all the papers and	1
3	What about the biggest misconception of	corpus	linguistics that you've encountered?	2
4	group of undergraduate students what they thought	corpus	linguistic was, guessed that it was the study of	3
5	more seriously, I think a common misconception of	corpus	linguistics is that it is just about using the to	3
6	om COCA or a list of collocations from any online	corpus	doesn't make you a corpus linguist or corpus rese	3
7	cations from any online corpus doesn't make you a	corpus	linguist or corpus researcher. Relating to the le	3
8	line corpus doesn't make you a corpus linguist or	corpus	researcher. Relating to the legal studies interdi	3
9	ne of my teachers, John Sinclair, who warned us a	corpus	is not a simple object, it's really easy to deriv	3
10	g to make up one just for you. What on earth does	corpus	linguistics have to do with jazz music?	4
11	in the school of music, Martin Norgaard. He had a	corpus	of jazz solos in which there were no words but th	5
12	us seven if somebody jumps up. With the help of a	corpus	analysis tool, basic analytic techniques – we use	5
13	, basic analytic techniques – we used AntConc – a	corpus	of, I think, about 450 solos from a database call	5

You can enable regular expressions to conduct more advanced searches (e.g., using *oh,? my god* to find all instances of “oh my god” as well as “oh, my god”). You can adjust the concordance by:

- Sorting the resulting concordance lines alphabetically by the right or left co-text/surrounding text, or by ascending order of the text ID (i.e., this means the line number of where the search term occurs in the text).
- Increasing (or decreasing) the amount of co-text/surrounding text.

In addition to viewing all relevant instances of the search term, you can download them as a .csv file for further analysis.

As mentioned above, this tool also allows you to analyse ‘dialogic’ structures in your data and/or analyse the meta-data that are associated with each occurrence of the search term – for instance, identifying whether a particular search term only occurs in the speech of a particular speaker. This feature of the tool is possible if you’ve set up your text data as a ‘structured’ .csv file where you have separate columns for text and other metadata attributes (e.g., speaker, date) or if you’ve set up the columns to indicate some kind of dialogic structure (e.g., one column includes the question and the next column includes the respective answer). This is explained and illustrated further in the File Upload section below, where an example of a structured .csv file is given.

There is also an additional option that allows users to analyse discourse structures in unstructured text (i.e., a text that does **not** contain different columns with aligned text pairs). Your text **does** need to contain a symbol that is consistently used to identify a structure – for

example, a question mark to identify a question. This is explained and illustrated further below.

It should be noted that this notebook is not meant to feature all types of analyses offered by current off-the-shelf concordancers and should instead be considered as complementary to such existing tools. You may want to use this tool if you are interested in using a concordancer for dialogic analysis or exploring the relationship between search terms and meta-data. Importantly, this notebook still requires further development but is a proof-of-concept example that users are able to test with small datasets.

## File upload


This notebook will allow you to upload a single txt file or csv file. If you want to analyse a corpus, you need to make sure you prepare your data accordingly before using the notebook.

1. Execute the cell:

```
[ ]: from ipywidgets import FileUpload
      from src.atap_widgets.concordance import ConcordanceLoader
      uploader = FileUpload(accept=".csv,.txt")
      display(uploader)
```


2. Click 'Upload (0)'. A window should appear prompting you to select a single txt file or csv file.

```
[1]: from ipywidgets import FileUpload
      from src.atap_widgets.concordance import ConcordanceLoader
      uploader = FileUpload(accept=".csv,.txt")
      display(uploader)
```

 Upload (0)

3. Click 'Open' after you've selected the file you want to upload.
4. The tool should start loading the selected file. Be aware that there is no progress indicator, but you will get a message if you run the next cell prior to the uploading process having completed.

```
[1]: from ipywidgets import FileUpload
      from src.atap_widgets.concordance import ConcordanceLoader
      uploader = FileUpload(accept=".csv,.txt")
      display(uploader)
```

 Upload (1)

Note: you can only upload one file. If you try to upload another file, the previous file you uploaded will be replaced by the new one.

## Concordancer – Structured data

You can now use the tool to retrieve a concordance for a search term. As a reminder, this concordancer is designed to help analyse dialogic structures (question-answer; post-response) and/or the metadata associated with your search term (such as the identity of the speaker,

the date, etc.). To do so, you must upload a ‘structured’ .csv file – a mocked-up example is provided below:

	A	B	C	D	E	F
1	text	speaker	affiliation	date	likes	response
2	In this column would be the text that the Concordancer searches for instances of the search term.	Chao	sydney informatics hub	13/10/2023	5	this would be the response to the post or question that is included in the text column A
3	here would be the text that was produced by the speaker in column B with the affiliation in column C	Monika	sydney corpus lab	13/10/2023	8	here you would have the response or comment to Monika's text
4	text by Jack is here, posted on the date noted in column D	Jack	sydney informatics hub	13/10/2023	3	this would be the relevant response to Jack's text
5	text by the speaker in column B and you can see how many people have liked this post in column E	Hamish	sydney informatics hub	13/10/2023	6	this would be the text for the response to Hamish's post

Note: you can save an .xlsx spreadsheet as .csv file within Excel via ‘Save as’. Make sure that the text you want to analyse is included in the column titled ‘text’.

This ‘structured data’ concordancer can be used even if you upload a plain text file that contains only text (unstructured data), but in this case you won’t be able to analyse dialogic structures or metadata.

1. Once you have uploaded your structured data, execute the cell:

```
[ ]: uploaded = len(uploaded.value) > 0
if uploaded:
    uploaded_file = uploaded.value[0]
    file_name = uploaded_file.name
    try:
        file_content = uploaded_file.content.tobytes().decode('utf-8')
    except UnicodeDecodeError:
        file_content = uploaded_file.content.tobytes().decode('latin-1')
    with open(file_name, "w") as fp:
        fp.write(file_content)

    file_type = uploaded_file.name[-3:]

    concordance_loader = ConcordanceLoader(path=file_name, type=file_type)
    concordance_loader.show()
else:
    print("Ensure you upload a file!")
```

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

Search term:

☐ Enable regular expressions
☒ Ignore case
☒ Match whole words

Page: 
Window size (characters):

Sort by:

Show More ...

Filename (without.xlsx extension)

If you didn’t upload a file, you will get a message saying, “Ensure you upload a file!”.

3. Toggle the boxes below the search field to enable or disable regular expression matching, case sensitivity, and whole word matching. By default, regular expression is disabled (i.e., box is unchecked) while case sensitivity and whole word matching are enabled (i.e., box is checked).
4. By default, the tool will display up to 50 characters of the co-text / surrounding text. You can change this by adjusting the number in the 'Window size' field. (Note: You can still adjust this after retrieving the concordance lines).
5. Use the 'Sort by' dropdown to sort by text\_id (i.e., this means the line number of where the search term occurs in the text), left context, or right context. (Note: You can still adjust this after retrieving the concordance lines).
6. Enter a search term into the search field and press enter on your keyboard to generate a concordance for the search term. The following example shows the concordance of the word *Canberra* in a sample text (prepared from a publicly available transcript of an interview with Australian Prime Minister Anthony Albanese). In this initial example no metadata are shown, because the metadata 'speaker' and 'role' were not selected:

Search term:

☐ Enable regular expressions
 ☒ Ignore case
 ☒ Match whole words

Page: 
 Window size (characters):

Sort by:

Show More ...

Filename (without.xlsx extension)

	left_context	match	right_context	text_id	
0	alia is Anthony Albanese and he's with us on this	Canberra	Day public holiday. Prime Minister, thanks for ma	0	
1	ay. Prime Minister, thanks for making time on ABC	Canberra	Breakfast.	0	
2	ing the balloons. And what an amazing setting for	Canberra.		1	
3		It is the best that	Canberra	has to offer, as locals know, on a morning like t	4
4	like today. I mean, what is a typical weekend in	Canberra	for you, given you have pledged and continue to l	4	

7. If your data contains metadata columns, use the 'Show More' field to select a metadata column to display. Otherwise, the only option available is "text". You can select more than one metadata category by using the left-click button on your mouse while holding the Ctrl or Command button. The following example shows the speaker metadata for each instance of the search term, showing users who said the search term in the uploaded sample text:



Search term: 

☐ Enable regular expressions
☒ Ignore case
☒ Match whole words

Page: 
Window size (characters):

Sort by: 

text\_id

Show More ... 

text
speaker
role

Filename (without.xlsx extension) 

Export to Excel

	left_context	match	right_context	text_id	speaker	
0	alia is Anthony Albanese and he's with us on this	Canberra	Day public holiday. Prime Minister, thanks for ma	0	ADAM SHIRLEY	
1	ay. Prime Minister, thanks for making time on ABC	Canberra	Breakfast.	0	ADAM SHIRLEY	
2	ing the balloons. And what an amazing setting for	Canberra.		1	ANTHONY ALBANESE	
3		It is the best that	Canberra	has to offer, as locals know, on a morning like t	4	ADAM SHIRLEY
4	like today. I mean, what is a typical weekend in	Canberra	for you, given you have pledged and continue to l	4	ADAM SHIRLEY	
5	and stunning weekend. My son, Nathan, was down in	Canberra	as well. So, we played tennis. We took advantage	5	ANTHONY ALBANESE	

8. As mentioned, the analysis also enables dialogic analysis if your uploaded .csv file consists of related text pairs, such as question-answer or social media post-response. The example below shows a concordance of the word *hurt* in a sample text (prepared from a publicly available transcript of an interview with former Australian Prime Minister Julia Gillard). The matching word is presented on the left, aligned with the full interview response ('text') and the relevant interview question that the text responds to ('question').

Show More ... 

text
question

Filename (without.xlsx extension) 

Export to Excel

	left_context	match	right_context	text_id	text	question
0	nowledge, we are not aware of any Australian being	hurt	. So we don't know of anybody at this stage. We do	4	To the best of our knowledge, we are not aware of any Australian being hurt. So we don't know of anybody at this stage. We do know though that there would still be people who are anxious, they know that they've got family or friends in Boston, who either were going in the marathon or going to watch the marathon, so we're saying please firstly try and contact your loved one and if you're not able to do that, then we do have a consular support line and we've got people working to ascertain everybody's safety. But we do not have any information before us at the moment which would lead us to believe that an Australian has been hurt.	As far as we know yet, were any Australians hurt in the bombing?
1	ld lead us to believe that an Australian has been	hurt	.	4	To the best of our knowledge, we are not aware of any Australian being hurt. So we don't know of anybody at this stage. We do know though that there would still be people who are anxious, they know that they've got family or friends in Boston, who either were going in the marathon or going to watch the marathon, so we're saying please firstly try and contact your loved one and if you're not able to do that, then we do have a consular support line and we've got people working to ascertain everybody's safety. But we do not have any information before us at the moment which would lead us to believe that an Australian has been hurt.	As far as we know yet, were any Australians hurt in the bombing?

9. The tool will only show up to 19 concordance lines on page 1. If the tool retrieved 20 or more concordance lines for your search term, then change the number shown next to "Page" (default is 1) to view more concordance lines.

10. To save a copy of the resulting concordance, enter a name for the file in the field for 'Filename'. In the example below, the filename is 'test'.

Search term:

☐ Enable regular expressions ☒ Ignore case ☒ Match whole words

Page:  Window size (characters):

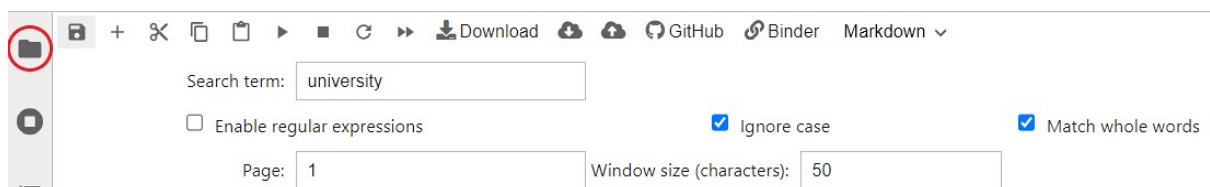
Sort by:

Show More ...

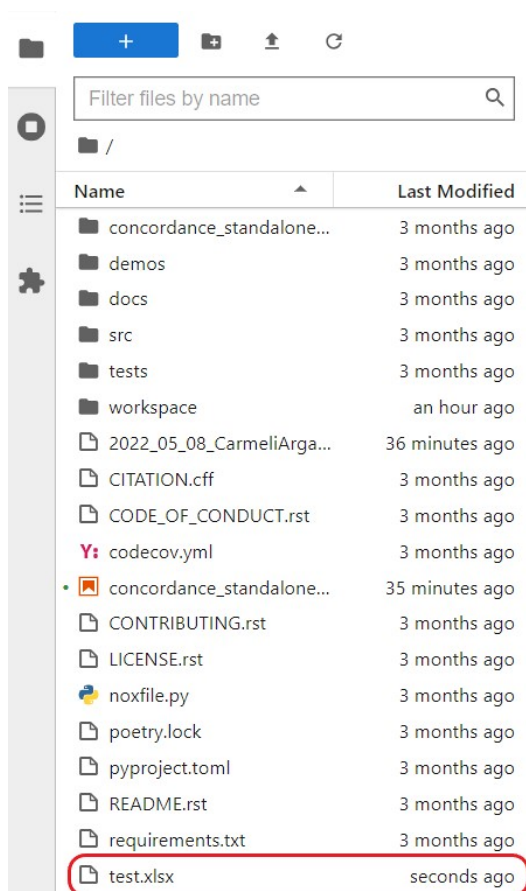
Filename (without.xlsx extension)  [Export to Excel](#)

11. Click the button labelled 'Export to Excel'.

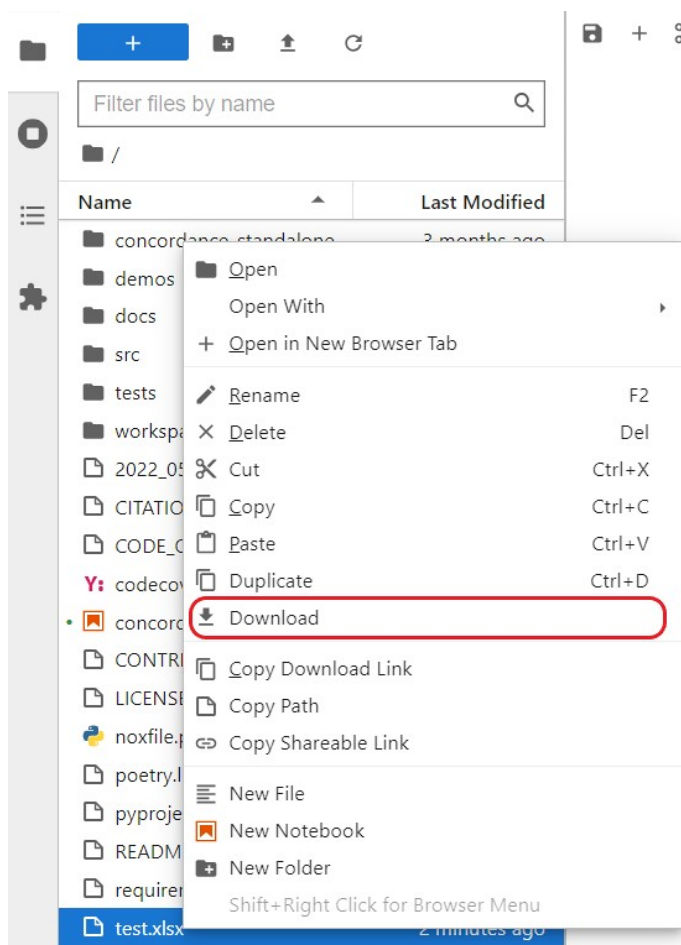
12. Open the file browser by clicking on the file symbol on the far-left menu (circled in red below):



13. You should see the xlsx file with the filename you specified in step 9. For example:



14. To download the spreadsheet, simply right click on the file and click 'Download':



## Concordancer – Unstructured data

This Concordancer also allows you to analyse discourse structures in unstructured text (i.e., a text that does **not** contain different columns with aligned text pairs or aligned text-metadata pairs). However, your text needs to contain a symbol that is consistently used to identify a structure. For example, your text might use the question mark symbol (?) only after an interviewer's question or your text might use the colon symbol (:) only after speaker names/labels (and before their respective dialogue), as shown in the example below. In this example, the colon only occurs following the speaker name, and it does not occur elsewhere. It can thus be used to identify the speaker:

```
ROBBIE LOVE: What is the biggest change in corpus linguistics that you've
noticed since the beginning of your career?
MONIKA BEDNAREK: I'd say just the incredible growth together with the
diversification into all the different areas of linguistics. For me, that's
the biggest change. I know there are other kind of qualitative changes. But
for me, that's the one that I would pick.
ROBBIE LOVE: Brilliant. Number two - what is the biggest misconception of
corpus linguistics you have encountered?
MONIKA BEDNAREK: A few years ago, I was in a meeting and someone referred to
corpus linguistics as "oh, isn't that just Google Ngrams?" So, I guess, for
me, that would be the conception that you don't have to have a very specific
kind of training and theoretical and methodological knowhow to do corpus
linguistics and also, that it is all about counting or quantitative
information where it does also include qualitative analysis.
```



1. Adjust the “splitter” variable in the following code cell (e.g., a colon, a question mark, etc). This tells the notebook which symbol should be used to identify a text structure. By default, the symbol that’s being treated as a splitter is a colon (i.e., splitter = “:”).
2. Once you’re happy with the splitter you’ve set, execute the cell:

```
[ ]: splitter = ":"

uploaded = len(uploader.value) > 0
if uploaded:
    uploaded_file = uploader.value[0]
    file_name = uploaded_file.name
    try:
        file_content = uploaded_file.content.tobytes().decode('utf-8')
    except UnicodeDecodeError:
        file_content = uploaded_file.content.tobytes().decode('latin-1')
    with open(file_name, "w") as fp:
        fp.write(file_content)

    file_type = uploaded_file.name[-3:]

    concordance_loader = ConcordanceLoader(path=file_name, type=file_type, re_symbol_txt=splitter)
    concordance_loader.show()
else:
    print("Ensure you upload a file!")
```

3. Once completed, you should get several widgets to adjust the settings for the concordance. The options for regular expression matching, case sensitivity, whole word matching, size of text displayed, sorting, etc are the same as for the Concordancer (structured data), as explained above. However, the difference lies in the ‘key’ and ‘text’ choices, which will be explained below.

The screenshot shows a web-based interface for a concordance tool. At the top, there is a 'Search term:' input field. Below it, there are three checkboxes: 'Enable regular expressions' (unchecked), 'Ignore case' (checked), and 'Match whole words' (checked). Further down, there are two input fields: 'Page:' with the value '1' and 'Window size (characters):' with the value '50'. Below these is a 'Sort by:' dropdown menu currently set to 'text\_id'. To the left of a larger dropdown menu is the text 'Show More ...'. The larger dropdown menu has two options: 'key' and 'text'. At the bottom, there is a label 'Filename (without.xlsx extension)' followed by an 'Enter filename' input field and a green 'Export to Excel' button.

If the file you uploaded does not contain the symbol that’s being set as the splitter, then you will get an error message. In this case, you can upload a new file that does contain the splitter symbol that you specified and then execute the cell again. Alternatively, you could replace the symbol in the cell to one that consistently identifies a dialogic structure in your uploaded file.

4. Enter a search term into the search field and press enter on your keyboard to generate a concordance for the search term. Use the ‘Show More’ field to show the text that precedes the splitter symbol (called the ‘key’). You can select both ‘key’ and ‘text’ options by using the left-click button on your mouse while holding the Ctrl or Command button. The following example shows the key in a sample text. In this case, the key shows the

speaker who uttered the relevant instance of the search term, because the colon occurs at the end of speaker names in the uploaded file:

Search term:

☐ Enable regular expressions ☒ Ignore case ☒ Match whole words

Page:  Window size (characters):

Sort by:

Show More ...

Filename (without.xlsx extension)

	left_context	match	right_context	text_id	key
0	What is the biggest change in	corpus	linguistics that you've noticed since the beginni	0	ROBBIE LOVE
1	Number two – what is the biggest misconception of	corpus	linguistics you have encountered?	2	ROBBIE LOVE
2	s ago, I was in a meeting and someone referred to	corpus	linguistics as "oh, isn't that just Google Ngrams	3	MONIKA BEDNAREK
3	and theoretical and methodological knowhow to do	corpus	linguistics and also, that it is all about counti	3	MONIKA BEDNAREK
4	o end on a rather optimistic note. How do you see	corpus	linguistics continuing to make an impact on the w	4	ROBBIE LOVE
5	I think it actually depends on how	corpus	linguistics is going to develop. The world is suc	5	MONIKA BEDNAREK

- The tool will only show up to 19 concordance lines. If the tool retrieved 20 or more concordance lines for your search term, then change the number shown next to "Page" (default page number is 1) to view more concordance lines.
- To download a copy of the resulting concordance, follow the same steps as explained above for the Concordancer (structured data).

## Citing/Referencing Notebook

Citation:

Bednarek, M., Mather, M., Maras, K., & Croser, H. (2023). ATAP Concordancer (version 0.5.4) [Computer software]. [https://github.com/Australian-Text-Analytics-Platform/atap\\_widgets](https://github.com/Australian-Text-Analytics-Platform/atap_widgets). DOI: <https://doi.org/10.5281/zenodo.10146967>.

If you are using this notebook in your research, please 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/atap\\_widgets](https://github.com/Australian-Text-Analytics-Platform/atap_widgets).*

## 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

This notebook only supports text files encoded in UTF-8 (recommended), ascii, or western “latin 1” (iso-8859-1). You can use tools such as [EncodeAnt](#) to convert your files into UTF-8 if necessary.

In addition, your text file should not contain all text in one single line, so do not remove existing line breaks from your text files. This Concordancer cannot process lines with 5000 or more characters.

This notebook is a proof-of-concept example and requires further development and testing.