

Semantic Tagger notebook – help pages

Introduction

The Semantic Tagger tool is a Jupyter notebook containing code that was adapted and developed (with permission) from the Python Multilingual Ucrel Semantic Analysis System ([PyMUSAS](#)). This was done 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 automatically processes text data, categorises and annotates/tags words or multi-word expressions (MWEs, e.g. *New South Wales*) based on their meaning class (e.g. location/place). Currently, the tool has been designed to annotate text data using Paul Rayson's Ucrel Semantic Analysis System ([USAS](#)) semantic tagset. The automatic semantic tagging helps bypass the time-consuming task of manually coding a large amount of text data.

There are currently two semantic tagger notebooks, one for English, and another one that supports multiple languages including Chinese, Italian, and Spanish.

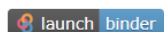
(Note: if you are unfamiliar with how to use Jupyter Notebooks, have a look at this [guide](#).)

Getting started

The tool is available on GitHub: the English semantic tagger is available [here](#) while the one for Chinese, Italian, and Spanish is available [here](#). You can launch the tool on Jupyter Notebook via Binder by clicking on one of the 'launch binder' buttons in the Setup section:

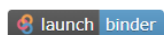
Setup

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 single sign-on (SSO) method with most (Australian or international) institutional login credentials or 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 is designed to assign semantic meaning to words or multi-word expressions (MWE) that occur in the files in your corpus. It allows you to preview the results (i.e., the first 500 tokens) for individual files in the form of a table – an example is shown in Figure 1 below. The preview table displays the tokens (i.e., words or punctuations) alongside other information such as their assigned USAS tags, definition of the assigned USAS tags, and the sentence where the word occurs. You can choose to preview the results for one text file or two sets of results (i.e., for two separate files) for comparison. As shown in Figure 1, the tool also displays the lemma (the form that would be listed in a dictionary) and part-of-speech (POS) tags for each word or MWE. This aspect of the tool may be of interest to those examining the association between the semantic class of a word (or MWE) and its part-of-speech function (word class, e.g. adjective, noun, etc) . The full results (i.e., table for all files) can be downloaded as a comma separated file (.csv) or as an excel spreadsheet (.xlsx) containing the tagged texts, or as a zipped folder (.zip) of tagged text (.txt) files, which can then be used to conduct further analysis.

	token	pos	usas tags	usas tags def	mwe	lemma	sentence
0	Facebook	PROPN	[Z3, Y2]	[Other proper names, Information technology and computing]	no	Facebook	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.
1	and	CCONJ	[Z5]	[Grammatical bin]	no	and	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.
2	Instagram	PROPN	[Z99]	[Unmatched]	no	Instagram	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.
3	,	PUNCT	[PUNCT]	[PUNCT]	no	,	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.
4	which	DET	[Z5]	[Grammatical bin]	no	which	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.

Figure 1. Screenshot of the resulting table showing the first 500 tokens (with only the first five visible) with their corresponding POS tag, USAS tag, etc.

The tool also produces another table that displays the frequency for every semantic category associated with each USAS tag. In essence, this table shows the number of tokens that have been assigned each semantic category/tag as seen in Figure 2.

	Other proper names	Information technology and computing	Grammatical bin	Unmatched	PUNCT	Getting and possession	Moving, coming and going	altogether	Time: Period	Speech acts	Personal names	Likely	Negative	Existing	Able/...
usas_tag	5.0	3.0	133.0	25.0	88.0	12.0	11.0	6.0	25.0	6.0	29.0	4.0	4.0	21.0	
pos	-	-	-	-	88.0	-	-	-	-	-	-	-	-	-	-
mwe	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Figure 2. Screenshot of part of the table showing all of the USAS tags and how many words were assigned certain tags

The tool also allows you to visualise the most commonly assigned semantic categories/tags or POS tags as well as the most frequent lemmas, tokens, or MWEs as bar graphs – an example is shown in Figure 3 below. It is possible to adjust the number of top items to display (i.e. in multiples of five). You can choose to visualise the most frequently assigned tags for the whole corpus or for individual files within the dataset. In the event you want to produce visualisations for individual files, you can choose to do this for just one file or for two separate files (e.g. for comparison). You can also choose to save all of the visualisations you produced (based on the set parameters) as jpg files.

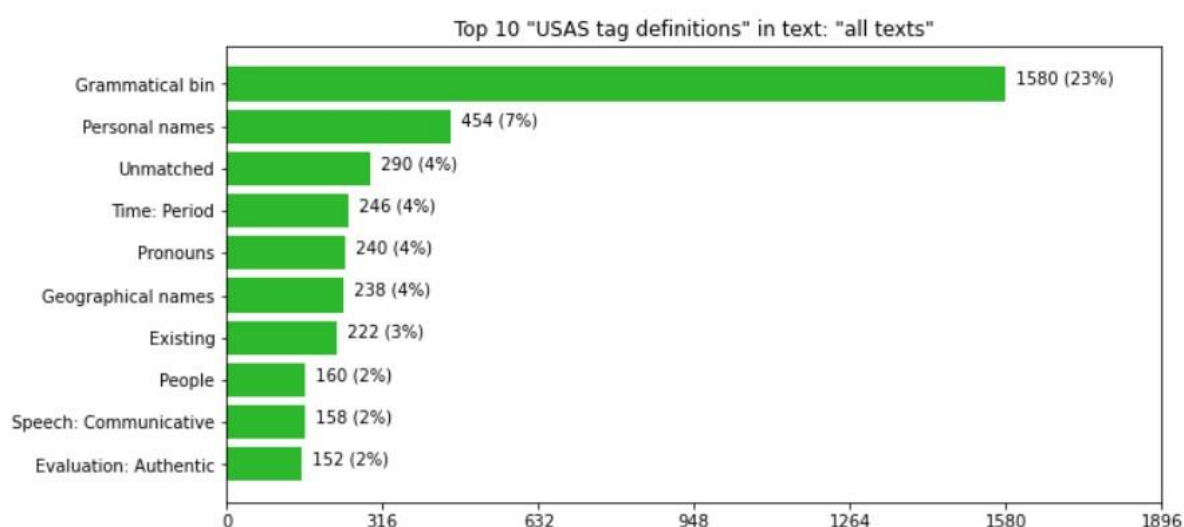


Figure 3. A bar graph showing the most frequently assigned USAS tags in across the entire corpus

In addition to [English](#), there is a [multilingual version](#) of the Semantic Tagger tool. This version of the tool can tag (traditional or simplified) Chinese (see Figure 2 below), Italian, or Spanish texts using the corresponding USAS tagset for these languages (with differing accuracy, coverage, etc.). (For those interested in Chinese, the tool can parse unsegmented Chinese text data and assign the appropriate semantic and POS tags.) The instructions below include information on how to use both the English and the multilingual version of the Semantic Tagger notebook.

Tagged text: chinese_text

The below table shows the first 500 tokens only. Use the above filter to show tokens with specific tags.

	token	pos	usas_tags	usas_tags_def	lemma	sentence
0	2012年	NOUN	[Z99]	[Unmatched]		2012年12月我在韩国留学的时候，有一天接到一个通知。
1	12月	NOUN	[Z99]	[Unmatched]		2012年12月我在韩国留学的时候，有一天接到一个通知。
2	我	PRON	[Z8]	[Pronouns]		2012年12月我在韩国留学的时候，有一天接到一个通知。
3	在	ADP	[Z5]	[Grammatical bin]		2012年12月我在韩国留学的时候，有一天接到一个通知。
4	韩国	PROP	[Z2]	[Geographical names]		2012年12月我在韩国留学的时候，有一天接到一个通知。
5	留学	VERB	[Z99]	[Unmatched]		2012年12月我在韩国留学的时候，有一天接到一个通知。
6	的	PART	[Z5]	[Grammatical bin]		2012年12月我在韩国留学的时候，有一天接到一个通知。
7	时候	NOUN	[T1.1]	[Time: General]		2012年12月我在韩国留学的时候，有一天接到一个通知。
8	,	PUNCT	[PUNCT]	[PUNCT]		2012年12月我在韩国留学的时候，有一天接到一个通知。
9	有一天	ADV	[Z99]	[Unmatched]		2012年12月我在韩国留学的时候，有一天接到一个通知。
10	接到	VERB	[A9+]	[Getting and possession]		2012年12月我在韩国留学的时候，有一天接到一个通知。

Figure 4. Screenshot of the resulting table (from the multilingual Semantic Tagger) showing Chinese words with their corresponding POS tag, USAS tag, etc.

Setup

Before you begin, you need to import the SemanticTagger and the necessary libraries and initiate them to run in this notebook.

1. Execute the cell to import the SemanticTagger and the necessary libraries:

```
[ ]: # import the SemanticTagger
print('Loading SemanticTagger...')
from semantic_tagger_en import SemanticTagger, DownloadFileLink

# initialize the SemanticTagger
st = SemanticTagger()
print('Finished loading.')
```

2. Once completed, you should get a message saying, “Finished loading”:

```
[1]: # import the SemanticTagger
print('Loading SemanticTagger...')
from semantic_tagger_en import SemanticTagger, DownloadFileLink

# initialize the SemanticTagger
st = SemanticTagger()
print('Finished loading.')
```

Loading SemanticTagger...

[nltk_data] Downloading package punkt to /home/jovyan/nltk_data...

[nltk_data] Unzipping tokenizers/punkt.zip.

Finished loading.

Loading the English SemanticTagger

1. Execute the cell:

```
[ ]: # select whether to include mwe extractions
st.loading_tagger_widget()
```

You should get a question asking you whether you want the tool to identify and tag Multi Word Expressions (MWEs), i.e., expressions formed by two or more words that behave like a unit such as 'South Australia':

```
[2]: # select whether to include mwe extractions
st.loading_tagger_widget()
```

[2]: Would you like to include multi-word expressions (mwe)?

☒ yes
☐ no

Warning: including mwe extraction will make the process much slower. For a corpus > 500 texts, we recommend choosing the non-MWE version.

Load Semantic Tagger

2. Select “yes” or “no”.
3. Click “Load Semantic Tagger”. You should get a message saying: “Loading the Semantic Tagger for the English language...”.
4. Once completed, you should get a message saying, “Finished loading”:

```
[2]: Semantic tagger without MWE extraction has been loaded and is ready for use.
```

☒ no

Load Semantic Tagger

Loading the Semantic Tagger for the English language...
 This may take a while...
 Finished loading.

If you’ve selected “no”, you will also see a message saying “Semantic tagger without MWE extraction has been loaded and is ready for use” (as shown in example above). If you’ve selected “yes”, you should instead see a message saying “Semantic tagger with MWE extraction has been loaded and is ready for use”.

Loading the **multilingual** SemanticTagger

1. Execute the cell:

```
[ ]: # select whether to include mwe extractions
st.loading_tagger_widget()
```

You should get several widgets:

```
[2]: # select whether to include mwe extractions
st.loading_tagger_widget()
```

[2]: **Select a language:**

spanish ▼

Would you like to include multi-word expressions (mwe)?

☒ yes
☐ no

Warning: including mwe extraction will make the process much slower. For a corpus > 500 texts, we recommend choosing the non-MWE version

Load Semantic Tagger

2. Select a language from the drop-down menu: Spanish, French, or Chinese.
3. Select “yes” or “no” for whether you want the tool to identify and tag Multi Word Expressions (MWE).

- Click “Load Semantic Tagger”. You should get a message saying: “Loading the Semantic Tagger for the selected language...”.
- Once completed, you should get a message saying, “Finished loading”:

```
[2]: # select whether to include mwe extractions
st.loading_tagger_widget()
```

[2]: Semantic tagger for chinese language has been loaded

chinese ▼

Semantic tagger without MWE extraction has been loaded and is ready for use.

☒ no

Load Semantic Tagger

Loading the Semantic Tagger for the selected language...
This may take a while...
Finished loading.

You will also get a message indicating which semantic tagger has been loaded. As shown in the example above, if you’ve selected Chinese, you will get “Semantic tagger for chinese language has been loaded”. You will also see a message indicating whether a semantic tagger with or without MWE extraction has been loaded.

Load the data


The notebook will allow you to upload text data in a text file (or a number of text files) or an excel spreadsheet. Please note that all text will be tagged. If your text contains any speaker names, transcription comments, or numbers – for example, timestamps – these will also be tagged. You may thus need to prepare your text before using the Semantic Tagger notebook.

- Execute the cell:

```
[ ]: # upload the text files and/or excel spreadsheets onto the system
display(st.upload_box)
print('Uploading large files may take a while. Please be patient...')
print('\033[1mPlease wait and do not press any buttons until the progress bar appears...\033[0m')
```

- Click ‘Upload your files’.

```
[3]: # upload the text files and/or excel spreadsheets onto the system
display(st.upload_box)
print('Uploading large files may take a while. Please be patient...')
print('\033[1mPlease wait and do not press any buttons until the progress bar appears...\033[0m')
```

 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...

A window should appear prompting you to select txt files, or a single csv file, xlsx file, or zip folder.

- Click ‘Open’ after you’ve selected the file(s) you want to upload.

- The tool should start loading the selected file(s). Once completed, you get a message saying, “Finished uploading files...” and another message stating the number of files that were uploaded (e.g., “100 text documents are loaded for tagging”):

```
[3]: # upload the text files and/or excel spreadsheets onto the system
display(st.upload_box)
print('Uploading large files may take a while. Please be patient...')
print('\033[1mPlease wait and do not press any buttons until the progress bar appears...\033[0m')
```

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

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, 47409.34it/s]
rm: cannot remove './input': No such file or directory
Pre-processing uploaded text...

100%|██████████| 100/100 [00:00<00:00, 35290.74it/s]
Finished uploading files.
100 text documents are loaded for tagging.

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

- Execute the following cell to view a snippet of the contents of the files you’ve uploaded:

```
[ ]: # display uploaded text
n=5

st.text_df.head(n)
```

By default, you can view (up to) five of the files. English text example:

```
[4]: # display uploaded text
n=5

st.text_df.head(n)
```

	text_name	text	text_id
0	2021_01_17_LaurenLancaster	Activists hold swim-in for trans inclusion at ...	dd181c3672dc3359353eda6497d585c3
1	2021_01_21_JulietteMarchant	NTEU appeals Tim Anderson Federal Court ruling...	cc8fe4275242db85d93392e5a705d4e0
2	2021_01_25_MaxShanahan	Health Minister rejects Invasion Day Covid pla...	29c99bc924bd2a1c4f744ff89f9d605b
3	2021_02_05_ShaniaOBrien	NUS condemns “horrific” assault on internation...	01aceb45444212877ad3c6b8a340ac85
4	2021_02_14_OliverPether	17 years on, activists continue to demand just...	26a7b1ff63231b9e8f181abf86827d40

Chinese text example:

```
[4]: # display uploaded text
n=5

st.text_df.head(n)
```

	text_name	text	text_id
0	chinese_text6	西班牙不願賦予古巴獨立，而古巴也不願作出任何讓步。科特柳的任期幾乎和麥金萊的任期一樣長，他在...	c3e3a86f246c5148374f518d7f3ca0eb
1	chinese_text	2012年12月我在韩国留学的时候，有一天接到一个通知。就是学校带我们留学生去旅行。地方是韩...	4964cba3ebdb10114c43faabb21ab141
2	chinese_text2	然而，這樣的處理也衍生了一些問題。自從2004年提出了興建人文大樓的構想，企業界陸續有人提供...	2552981f7169dd0545c5dda5d8e8156a
3	chinese_text3	那个旅社位于狭窄的胡同，好不容易找。不过我们足足找了半个小时，才找到。他们的热情让我很感动，...	a3bc7a7dc6325223d544683e1903db2c
4	chinese_text4	同時贏得了聲望和惡名。查爾斯·克拉克在其同年出版的專著《蘇門答臘島與西馬來西亞的豬籠草》中拒...	5ec408be0143c41453d706888d6e8baf

You can adjust this by changing the number value in “n=5” code (to e.g., n=2, n=10, n=20) before you execute the cell.

Apply semantic tagging

Once your texts have been uploaded, they are ready to be semantically tagged. Depending on the size of the uploaded corpus, this process may take a long time. Once it's completed, you can then preview/display the tagged text in a table format and download the results to your computer. You can choose to display one tagged text file or compare two tagged text files with each other.

1. Execute the cell to start tagging your corpus/dataset:

```
[ ]: # add semantic taggers to the uploaded texts
print('Processing and adding semantic tags to your texts.')
print('The counter will start soon. Please be patient...')
st.tag_text()
```

2. Once completed, you should get a progress bar displaying 100%, similar to the one shown below:

```
[5]: # add semantic taggers to the uploaded texts
print('Processing and adding semantic tags to your texts.')
print('The counter will start soon. Please be patient...')
st.tag_text()
```

```
Processing and adding semantic tags to your texts.
The counter will start soon. Please be patient...
100%|██████████| 100/100 [00:33<00:00, 3.01it/s]
```

Preview/display tagged texts

1. Execute the cell:

```
[ ]: # display tagged text
st.display_two_tag_texts()
```

2. Once completed, you should get several widgets to adjust the settings for the preview of one or two tagged texts:

The image shows two identical sets of interactive widgets side-by-side. Each set consists of:

- A 'Select text:' dropdown menu with the placeholder text 'Choose text to display...'.
- A 'pos:' dropdown menu with 'all' selected.
- A 'usas tag:' dropdown menu with 'all' selected.
- A 'Display tagged text' button located below the dropdowns.

3. Select which text you want to preview.
4. Select which part-of-speech (pos) tag to view. By default, “all” part-of-speech tags will be displayed. (Note: Other pos tag options will be available once you’ve completed step 6.)
5. Select which USAS tag to view. By default, “all” USAS tags will be displayed. (Note: Other USAS tag options will be available once you’ve completed step 6.)

- Once you're happy with the settings, click "Display tagged text". You will get a table displaying the first 500 tagged tokens/words along with information such as the assigned pos and USAS tags. English text example:

[6]: Select text: Select text:

pos:
 ADJ
 ADP
 ADV
 AUX

usas tag:
 Allowed
 Anatomy and physiol
 Architecture, houses a

Display tagged text

pos:
 usas tag:

Display tagged text

Tagged text: 2021_01_17_LaurenLancaster
 The below table shows the first 500 tokens only. Use the above filter to show tokens with specific tags.

	token	pos	usas_tags	usas_tags_def	lemma	sentence
0	Activists	NOUN	[G1.2, S2]	[Politics, People]	activist	Activists hold swim-in for trans inclusion at McIver's Baths\rThe transphobic entry policy of the baths has been met with strong public backlash.
1	hold	VERB	[M2]	[Putting, pulling, pushing, transporting]	hold	Activists hold swim-in for trans inclusion at McIver's Baths\rThe transphobic entry policy of the baths has been met with strong public backlash.
2	swim	NOUN	[M4]	[Sailing, swimming, etc.]	swim	Activists hold swim-in for trans inclusion at McIver's Baths\rThe transphobic entry policy of the baths has been met with strong public backlash.
3	-	PUNCT	[PUNCT]	[PUNCT]	-	Activists hold swim-in for trans inclusion at McIver's Baths\rThe transphobic entry policy of the baths has been met with strong public backlash.
4	in	NOUN	[M6]	[Location and direction]	in	Activists hold swim-in for trans inclusion at McIver's Baths\rThe transphobic entry policy of the baths has been met with strong public backlash.

Chinese text example:

[6]: Select text: Select text:

pos:
 ADJ
 ADP
 ADV
 CCONJ

usas tag:
 Able/intelligent
 Alive
 Allowed

Display tagged text

pos:
 usas tag:

Display tagged text

Tagged text: chinese_text
 The below table shows the first 500 tokens only. Use the above filter to show tokens with specific tags.

	token	pos	usas_tags	usas_tags_def	lemma	sentence
0	2012年	NOUN	[Z99]	[Unmatched]		2012年12月我在韩国留学的时候，有一天接到一个通知。
1	12月	NOUN	[Z99]	[Unmatched]		2012年12月我在韩国留学的时候，有一天接到一个通知。
2	我	PRON	[Z8]	[Pronouns]		2012年12月我在韩国留学的时候，有一天接到一个通知。
3	在	ADP	[Z5]	[Grammatical bin]		2012年12月我在韩国留学的时候，有一天接到一个通知。
4	韩国	PROP	[Z2]	[Geographical names]		2012年12月我在韩国留学的时候，有一天接到一个通知。
5	留学	VERB	[Z99]	[Unmatched]		2012年12月我在韩国留学的时候，有一天接到一个通知。
6	的	PART	[Z5]	[Grammatical bin]		2012年12月我在韩国留学的时候，有一天接到一个通知。
7	时候	NOUN	[T1.1]	[Time: General]		2012年12月我在韩国留学的时候，有一天接到一个通知。
8	,	PUNCT	[PUNCT]	[PUNCT]		2012年12月我在韩国留学的时候，有一天接到一个通知。
9	有一天	ADV	[Z99]	[Unmatched]		2012年12月我在韩国留学的时候，有一天接到一个通知。

You will also get another table showing how many words were assigned each pos and USAS tags. For example:

	Politics	People	Putting, pulling, pushing, transporting	Sailing, swimming, etc.	PUNCT	Location and direction	Grammatical bin	Business: Selling	Inclusion	Unmatched	Furniture and household fittings	Cleaning and personal care	Places	Wanted	p
usas_tag	5.0	28.0	1.0	9.0	116	18.0	253.0	16.0	4.0	52.0	15.0	15.0	7.0	10.0	
pos	-	-	-	-	116	-	-	-	-	-	-	-	-	-	

7. You should now be able to choose different pos and usas tag options to display specific words or groups of words (e.g., all words tagged as ‘ADJ’ and “Able/intelligent”, or those tagged as “NOUN” and “Location and direction”). You can select more than one pos and/or usas tag by using the left-click button on your mouse while holding the Ctrl or Command button. Once you’re happy with your selection, click “Display tagged texts” and you should get a new table. For example, the following table displays all words tagged as different sub-categories of “Evaluation”:

Select text: 2021_01_17_LaurenLancaster

pos: all
ADJ
ADP
ADV
AUX

usas tag: Ethical
Evaluation: Bad
Evaluation: False
Evaluation: Good
Evaluation: Particular

Display tagged text

Select text: Choose text to display...

pos: all

usas tag: all

Display tagged text

Tagged text: 2021_01_17_LaurenLancaster

	token	pos	usas_tags	usas_tags_def	lemma	sentence
465	bad	ADJ	[A5.1-]	[Evaluation: Bad]	bad	\rShe reflected that "ambiguous policy leaves [the] door open to discrimination," which is fundamentally a bad outcome.
26	backlash	NOUN	[A5.1-]	[Evaluation: Bad]	backlash	Activists hold swim-in for trans inclusion at McIver's Baths\rThe transphobic entry policy of the baths has been met with strong public backlash .
514	fit	VERB	[N3.2, A5.1+]	[Measurement: Size, Evaluation: Good]	fit	She concluded, with vocal support from participants, that "this is not about keeping women safe, this is about policing women's bodies... we must ensure that all women are welcome to use the baths... [not just those] who fit [an inappropriate, narrowly defined] idea of womanhood."
407	misconception	NOUN	[X4.1, A5.2-]	[Mental object: Conceptual object, Evaluation: False]	misconception	\rAnderson continued with an analysis of the Baths ban, first highlighting the misconception that "all trans women want gender affirming surgery" and second challenging the notion that bottom surgery is a necessary means of confirming gender identity, labelling it as "plain outdated."

8. If you want to view another tagged text, adjust the settings in the second set of widgets by repeating these steps. To choose a different text, first clear the currently selected text (e.g. by selecting/highlighting it and pressing “Delete” or “Backspace”). This enables you to select a different text for the preview. (Tip: typing part of the targeting text name will filter the dropdown list for a quicker selection.)

Analyse tagged texts

You can also analyse the tagged texts through simple visualisations.

1. Execute the cell:

```
[ ]: # analyse tagged texts
st.analyse_two_tags()
```

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

[7]:

Select text: Choose text to display...

Select entity to show:

usas_tags_def

Select tag/lemma/token:

None

Select n:

5

Show top entities

Save analysis

Select text: Choose text to display...

Select entity to show:

usas_tags_def

Select tag/lemma/token:

None

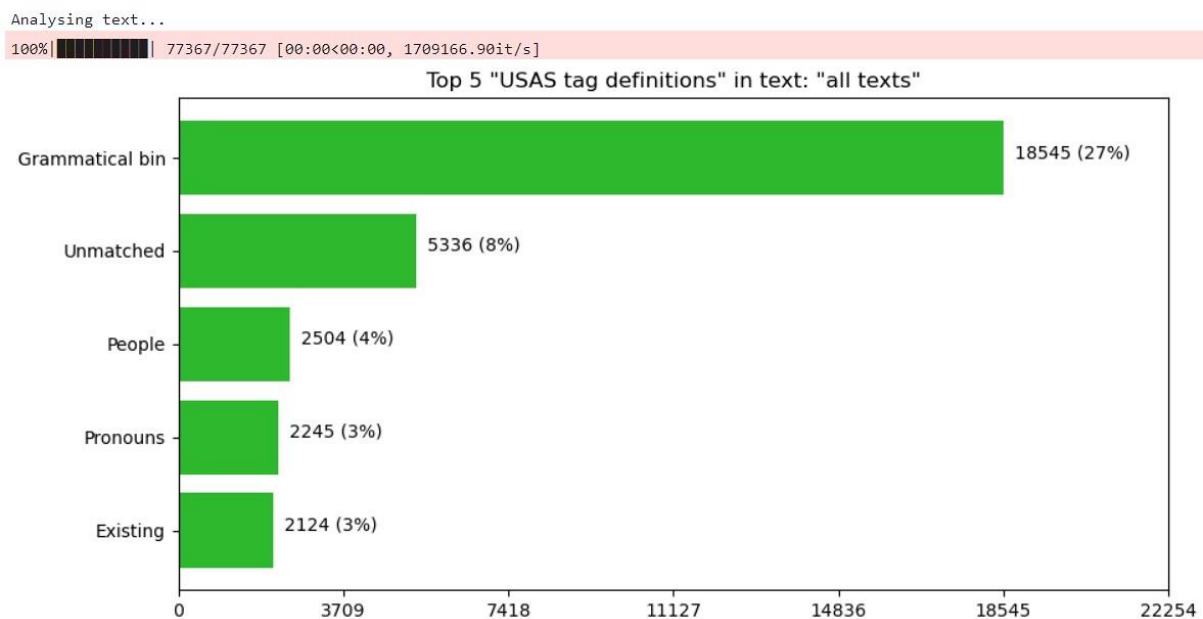
Select n:

5

Show top entities

Save analysis

3. Select the text you want to include in the visualisation.
4. Select which entities you want to include in the visualisation: "usas_tags_def" (i.e., USAS tags), "pos", "lemma", or "token".
5. Select n to display the top number of entities or words. By default, the tool will only display the top five entities (i.e., Select n: 5) in the entity category you've selected in step 4.
6. Once you're happy with the settings, click "Show top entities". You will get a bar graph displaying the top number of entity (i.e., USAS tags, POS tags, lemmas, or tokens) within the text you've selected. For example:



7. Repeat these steps to generate a new bar graph for a different text, different entity, or different number of entities. Each new bar graph you generate will replace the previous one generated. However, the software will keep a copy of all of the bar graphs generated for you to save later – this will need to be done before you shut down the kernel.
 Note: To choose a different text, first clear the currently selected text (e.g. by selecting/highlighting it and pressing "Delete" or "Backspace"). This enables you to select a different text for the preview. (Tip: typing part of the targeting text name will filter the dropdown list for a quicker selection.)

8. You can also view the top number of words within a subcategory of the type of entity you've selected in step 4. (Note: the number of subcategories you can choose from depends on the number you've set in step 5).

If you've selected "usas_tags_def", select the subcategory from the list under "Select USAS tag definition":

[7]: **Select text:**

Select entity to show: <div><input type="text" value="usas_tags_def"/> ▼</div> Select n: <div><input type="text" value="10"/></div> <div>Show top entities</div> <div>Save analysis</div>	Select USAS tag definition: <div><div>Grammatical bin Unmatched People Pronouns</div></div> Select n: <div><input type="text" value="5"/></div> <div>Show top words</div>
---	--

If you've selected "pos", select the subcategory from "Select pos tag definition":

[7]: **Select text:**

Select entity to show: <div><input type="text" value="pos"/> ▼</div> Select n: <div><input type="text" value="10"/></div> <div>Show top entities</div> <div>Save analysis</div>	Select Part-of-Speech Tag: <div><div>NOUN ADP PROPN VERB DET</div></div> Select n: <div><input type="text" value="5"/></div> <div>Show top words</div>
---	---

If you've selected "lemma", select the subcategory from "Select lemma definition":

[7]: **Select text:**

Select entity to show:

lemma ▼

Select n:

10

Show top entities

Save analysis

Select lemma:

the

be

to

of

Select n:

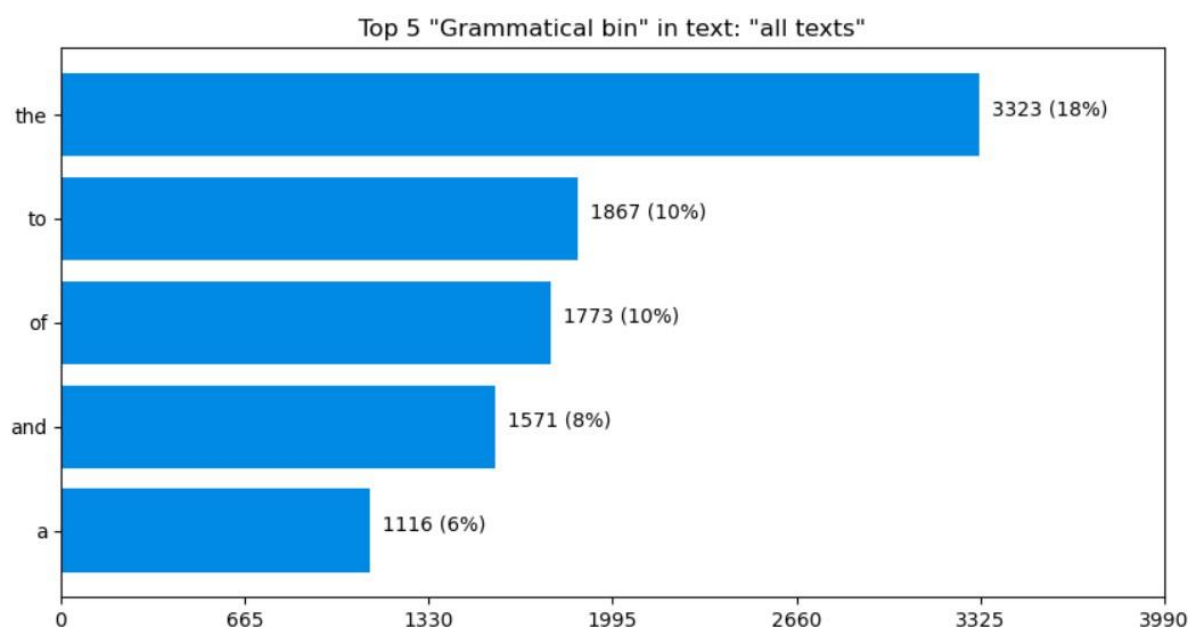
5

Show top words

No options are available if you've selected "tokens".

You can select more than one subcategory by using the left-click button on your mouse while holding the Ctrl or Command button. The tool will produce a separate graph for each subcategory you select.

9. Select to display the top number of words within the subcategory of entities you've selected (in step 8).
10. Once you're happy with the settings, click "Show top words". You will get a bar graph displaying the top number of words – i.e., most frequent – within the subcategory of entity you've selected. For example, the screenshot below shows that *the*, *to*, *of*, *and*, *a* are the top five most frequent words semantically tagged as the category "Grammatical bin":



11. Repeat these steps to generate a new bar graph for a different subcategory of the selected entity (i.e., USAS tag, pos tag, or lemma), or different number of entities. You can also

visualise the results of another tagged text by repeating the steps explained above (selecting a different text or corpus).

12. Click “Save analysis” to save all of the bar graph you produced. You should get a separate link to jpg files of each bar graph you generated:

[7]: **Select text:**

Select entity to show:

▼

Select n:

Show top entities

Save analysis

Select lemma:

be
of
to
and

Select n:

Show top words

Analysis saved! Click below to download:

- [Top-5-USAS-tag-definitions-in-text:-all-texts.jpg](#)
- [Top-5-Part-of-Speech-Tags-in-text:-all-texts.jpg](#)
- [Top-5-Part-of-Speech-Tags-in-text:-all-texts.jpg](#)
- [Top-5-lemmas-in-text:-all-texts.jpg](#)
- [Top-5-tokens-in-text:-all-texts.jpg](#)
- [Top-10-USAS-tag-definitions-in-text:-all-texts.jpg](#)
- [Top-5-Grammatical-bin-in-text:-all-texts.jpg](#)
- [Top-10-Part-of-Speech-Tags-in-text:-all-texts.jpg](#)
- [Top-5-NOUN-in-text:-all-texts.jpg](#)
- [Top-10-lemmas-in-text:-all-texts.jpg](#)
- [Top-4-the-in-text:-all-texts.jpg](#)

13. Click on each link to download the corresponding jpg of the bar graph.

Save tagged texts

You can save the tagged texts into a comma separated file (.csv) or an excel spreadsheet (.xlsx) containing the tagged texts, or as a zipped file of pseudo-xml tagged text files (.txt). You can then download them to your local computer and use them for further analysis if you wish.

1. Execute the cell:

```
[ ]: # save tagged texts
st.save_options()
```

2. Select which file type you want to save the results as: excel, csv, or pseudo-xml.

```
[8]: # save tagged texts  
st.save_options()
```

[8]: Select saving file type:

excel ▼

Save tagged texts

3. Click on “Save tagged texts”. You should get a link called “tagged_texts” with the file extension corresponding to the file type you’ve selected (i.e., xlsx for “excel”, csv for “csv”, or zip for “pseudo-xml”). For example:

```
[8]: # save tagged texts  
st.save_options()
```

[8]: Select saving file type:

excel ▼

Save tagged texts

Saving tagged texts in progress. Please be patient...

Tagged texts saved. Click below to download:

[tagged_texts.xlsx](#)

4. Click on the link to the file to save it to your computer.

Note: There may be an empty code cell at the end of the notebook, which you can ignore.

Citing/Referencing Notebook

Citation: Jufri, S. & Sun, C. (2022). Semantic Tagger. (version 1.0) [Jupyter notebook]. Australian Text Analytics Platform. <https://github.com/Australian-Text-Analytics-Platform/semantic-tagger>

You can adjust the year, version number and URL 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/semantic-tagger>.

In addition, please inform [ATAP](#) 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.