Phrase Tagging with Doccano

Doccano is an open-source data labeling tool intended for machine learning practitioners. It allows you to tag words and phrases in texts with a very easy-to-use drag and select user interface.

doccano-ux.webm

You can use Doccano to tag phrases and import your tags into phrase layers in LaBB-CAT. The broad steps of the process are:

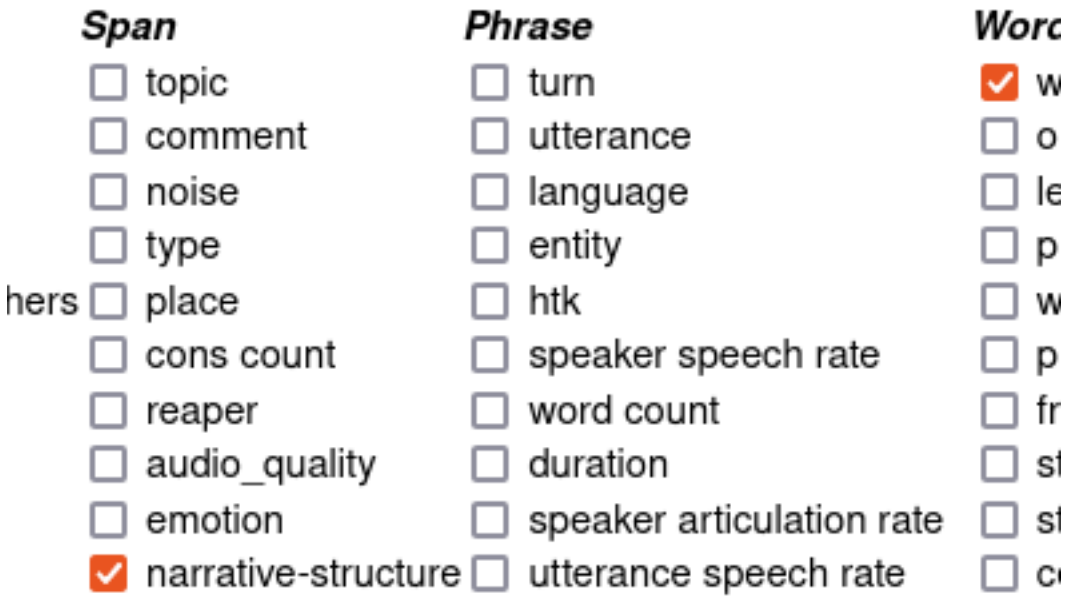
1. Export a selected set of transcripts from LaBB-CAT to the Doccano JSONL format
2. Import the resulting file into Doccano
3. Tag the texts as desired
4. Export the tagged texts from Doccano to a JSONL file
5. Import the tagged JSONL file into LaBB-CAT

> **Installation of the Doccano Formatter**
>
> LaBB-CAT uses 'formatter' modules to import and export files in the formats of different annotation tools.
> If you find that 'Doccano JSONL Dataset' is not an option for exporting or importing Doccano files, it's because the Doccano module is not installed on your LaBB-CAT instance.
> To install the Doccano formatter:
>
> 1. Download the format conversion module here: https://github.com/nzilbb/ag/blob/main/bin/nzilbb.formatter.doccano.jar
> 2. In LaBB-CAT, select the *converters* menu option
> 3. Press *Choose File* at the bottom, and select the nzilbb.formatter.doccano.jar file you downloaded in 1.
> 4. Press *Upload*
> 5. Press *Install*
>
> Doccano JSONL Dataset should now be an option for selection on the transcripts page.

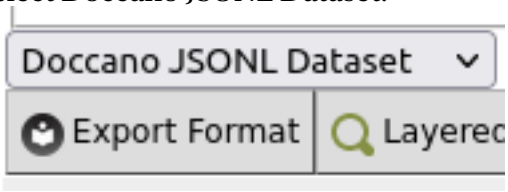
1. Export from LaBB-CAT

1. In LaBB-CAT, open the *transcripts* page.
2. Use the filters at the top to narrow the list down to the transcripts you want to export, and/or tick the target transcripts.

3. Click *Export Format*.
   A list of layers will appear, with a list of formats below.
4. If you wish to include any existing phrase/span layers in Doccano, tick the corresponding layers in the list.



   NB: Annotations on the layers you select here will be displayed in Doccano but cannot be edited; any changes to these annotations will be ignored when re-importing the dataset into LaBB-CAT.
5. Below the list of layers, there's a dropdown list of export formats.
   Select Doccano JSONL Dataset.[1]



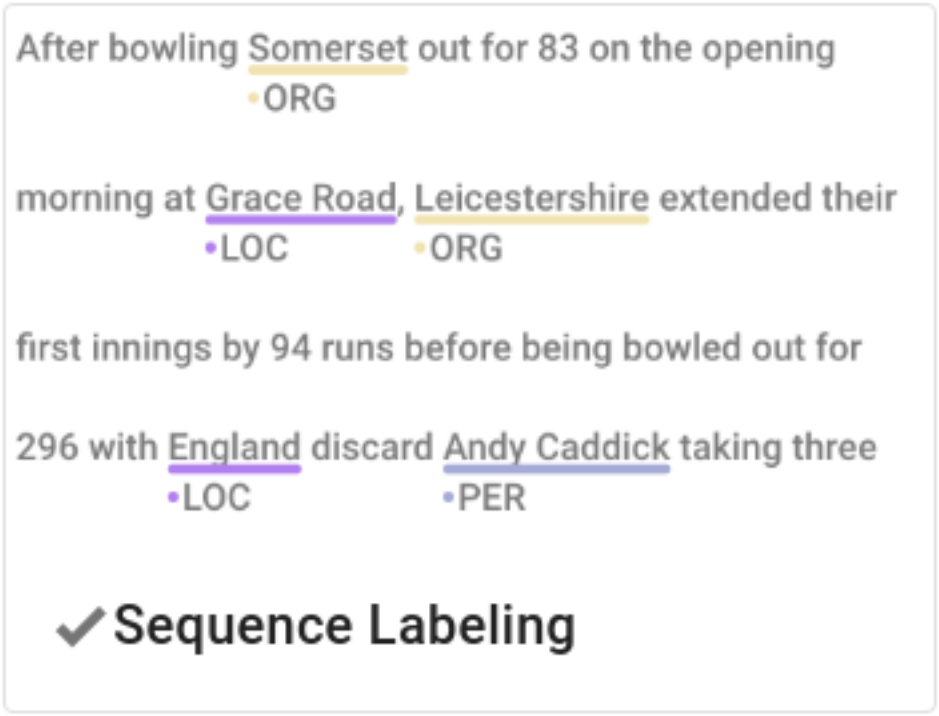
6. Click *Export Format*
7. Save the resulting …*jsonl* file.

2. Import into Doccano

1. In Doccano, you will need to create a project to import your texts into. Click *Projects* on the top right.
2. Press *Create* on the top left.

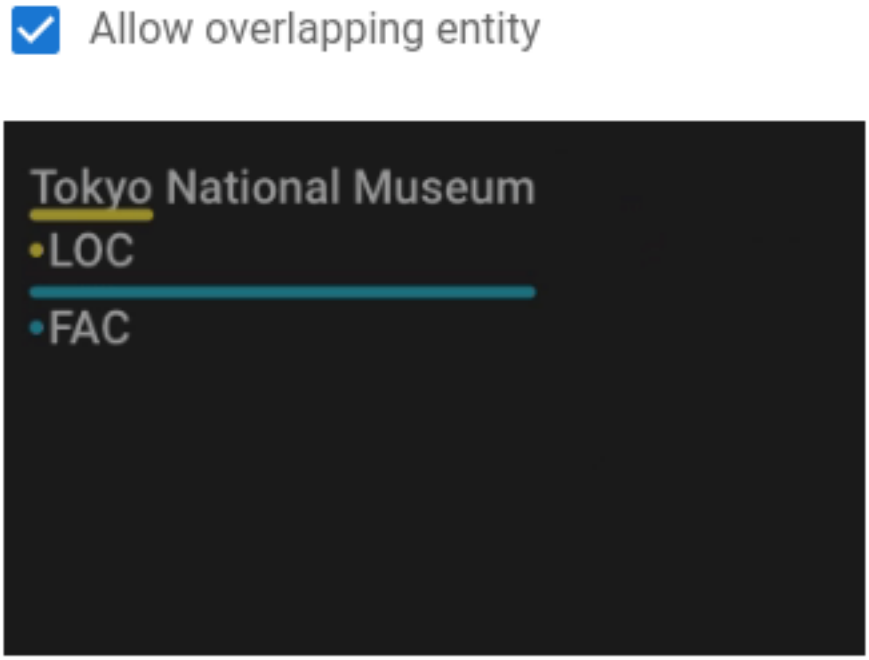
[1] If Doccano JSONL Dataset is not an option, then you need to install the Doccano JSONL Dataset formatter - see instructions on Installation of the Doccano Formatter

3. Select the *Sequence Labelling* option.



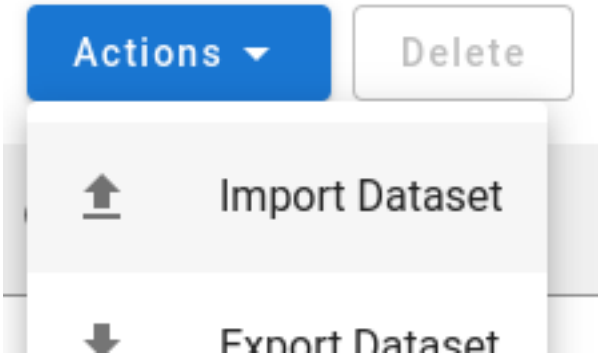
Sequence Labelling option in Doccano is ticked

4. Enter a name and description for your project.
5. Tick the *Allow overlapping entity* option.



6. Press *Create* at the bottom.
   This will create the project and take you to its Home page.
7. Press *Dataset* on the left.

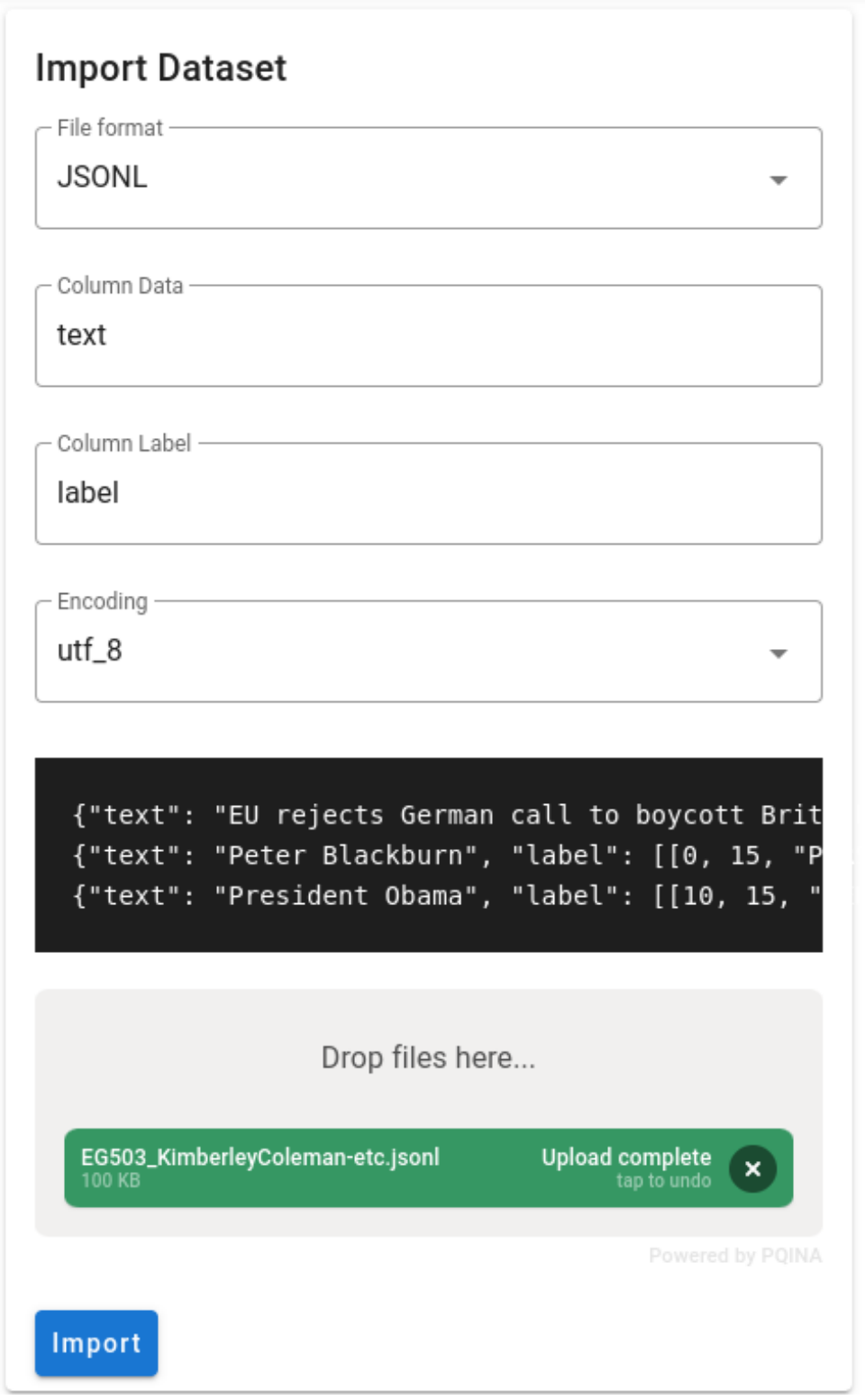
8. Move the mouse over the *Actions* button at the top, and select the *Import Dataset* option.



9. Select *JSONL* as the *File format*.

You can leave the other options that appear with their default values.

10. Find the *…jsonl* file you exported from LaBB-CAT earlier, and drap/drop it on to the grey area labelled *Drop files here…*
    (Alternatively you can click on the *Drop files here…* area, and find/select the *…jsonl* file.)


Import Dataset
File format
JSONL
Column Data
text
Column Label
label
Encoding
utf_8
{"text": "EU rejects German call to boycott Brit
{"text": "Peter Blackburn", "label": [[0, 15, "P
{"text": "President Obama", "label": [[10, 15, "
Drop files here...
EG503_KimberleyColeman-etc.jsonl
100 KB
Upload complete
tap to undo
Powered by PQINA
Import

11. Press *Import.*

Once the import is complete, you will see a list of texts on the *Dataset* page. The *Metadata* column will be full of text and numbers - this is normal; LaBB-CAT includes information in the *Metadata* that it needs to import the text back into LaBB-CAT correctly.

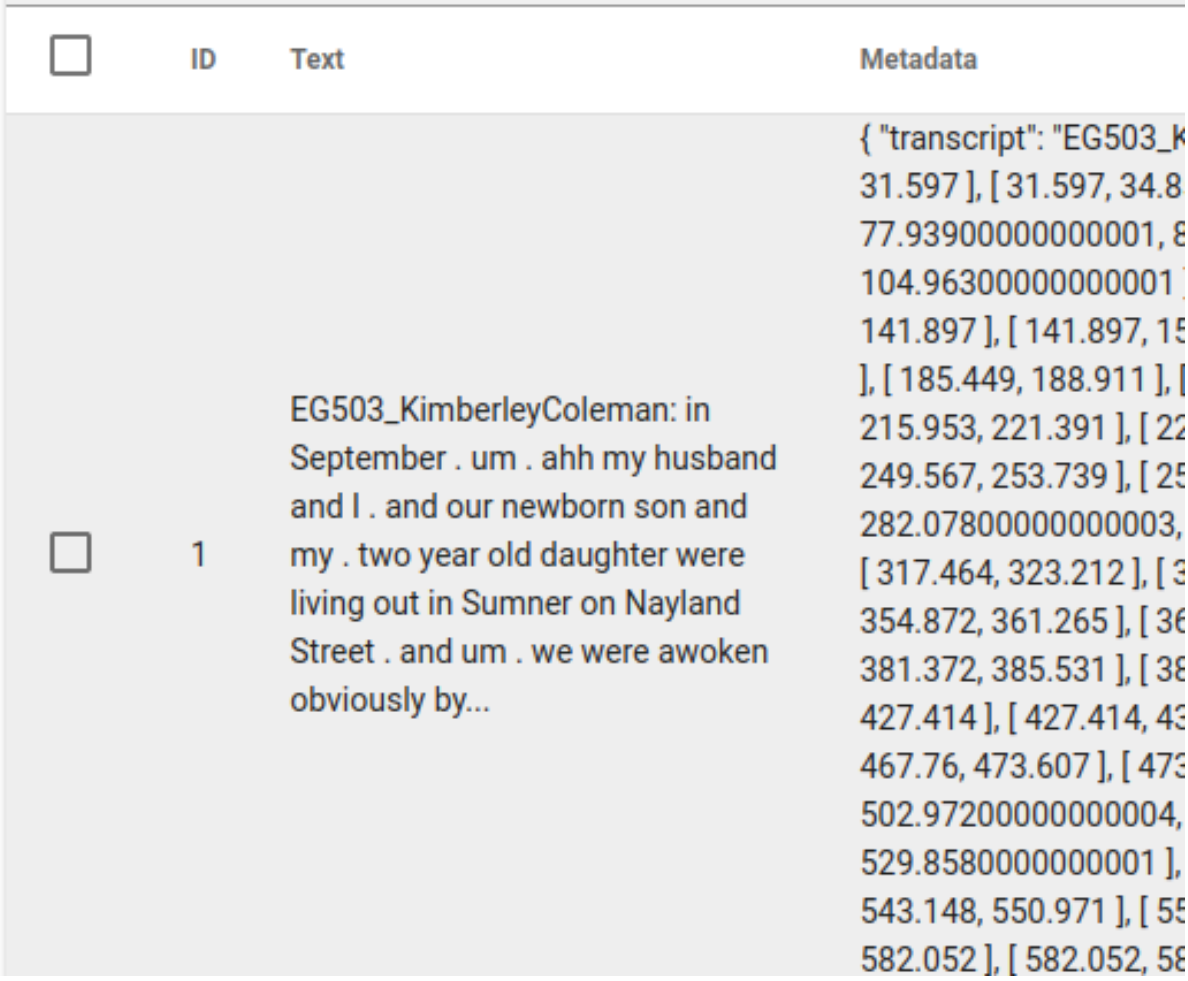


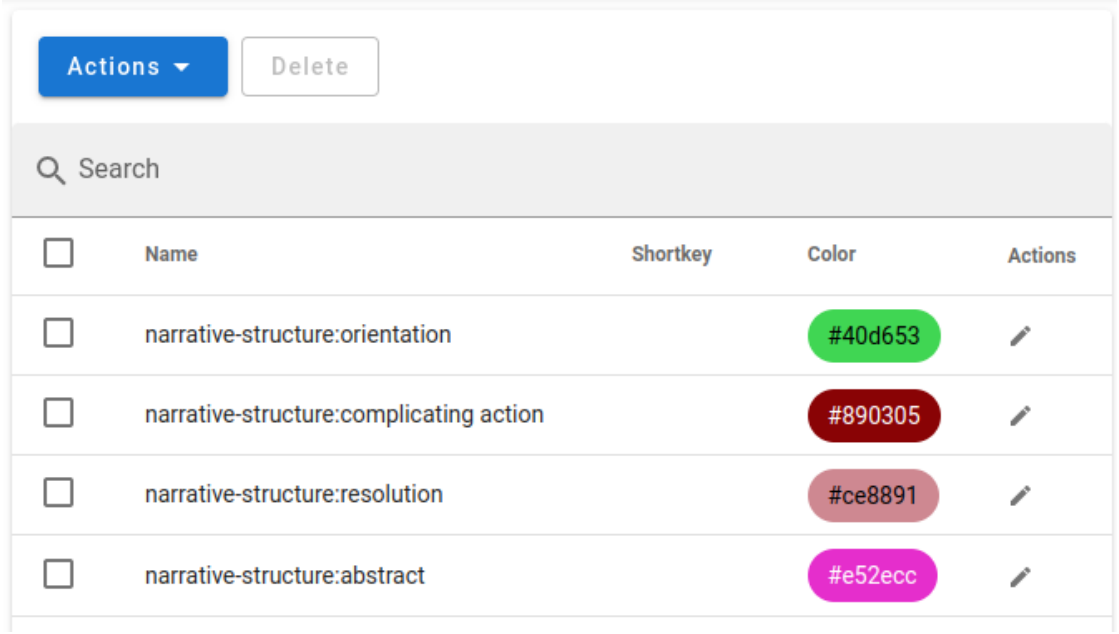
Figure 1: The imported dataset includes metadata to aid re-import into LaBB-CAT

3. Tag the texts

Before adding annotations to the texts, you need to create *Labels* in Doccano. These are the annotations you'll be able to add to words/phrases in the texts.

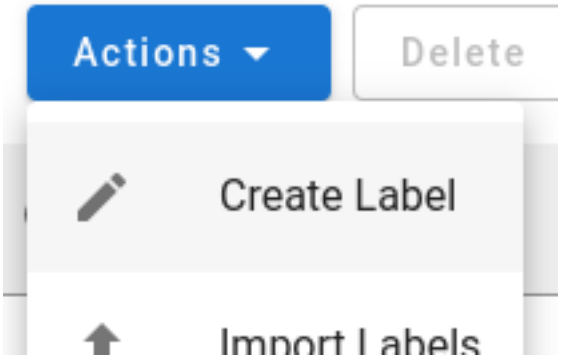
1. In Doccano, click *Labels* on the left-hand menu.
   If you exported any additional phrase/span layers from LaBB-CAT, you will see labels

for the resulting annotations already listed here.

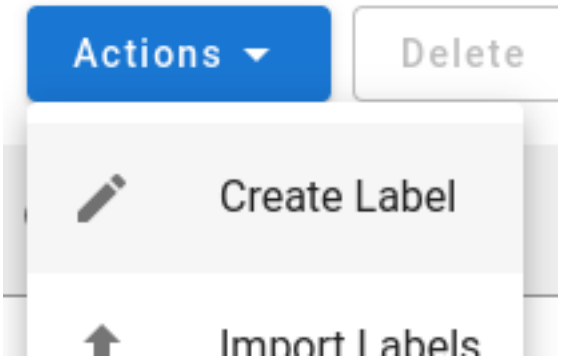


Each imported label is prefixed with the ID of the LaBB-CAT layer it came from, followed by a colon. This is the pattern you must follow with the labels you create.

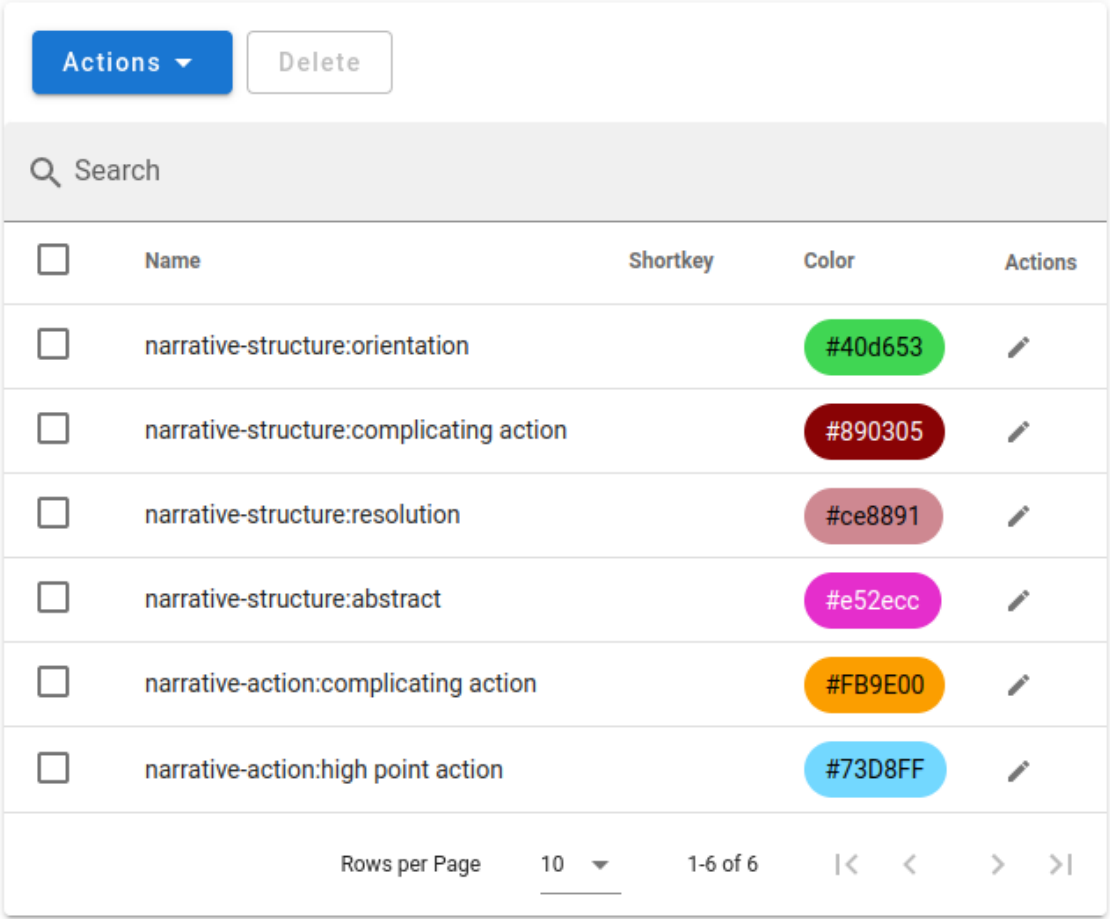
2. Move the mouse over the *Actions* button at the top and click the *Create Label* option.



3. Enter the label. This should be using the format:
`{LaBB-CAT-Layer-ID}:{LaBB-CAT-Label}`
e.g if you intend for your new annotations to be added to a LaBB-CAT layer called "narrative-action" and one of the possible labels in LaBB-CAT will be "complicating action", then the Label you create in Doccano should be: `narrative-action:complicating action`



4. Pick a colour for the label if you wish.
5. Assuming you want to add more than one label, click *Save and add another*.
6. Repeat the above steps for each label you would like to annotate with.
7. Once you've finished, click the *Labels* option on the left-hand menu.
   You should see the label's you've added, listed after the imported ones.



Annotating Texts

Now that you've configured the labels you're going to use, you can annotate the texts you imported:

In Doccano click *Start Annotation* at the top left.

(Alternatively, you can click *Dataset* and then press the *Annotate* button on a text of your choice)

You will see one of the texts you imported.

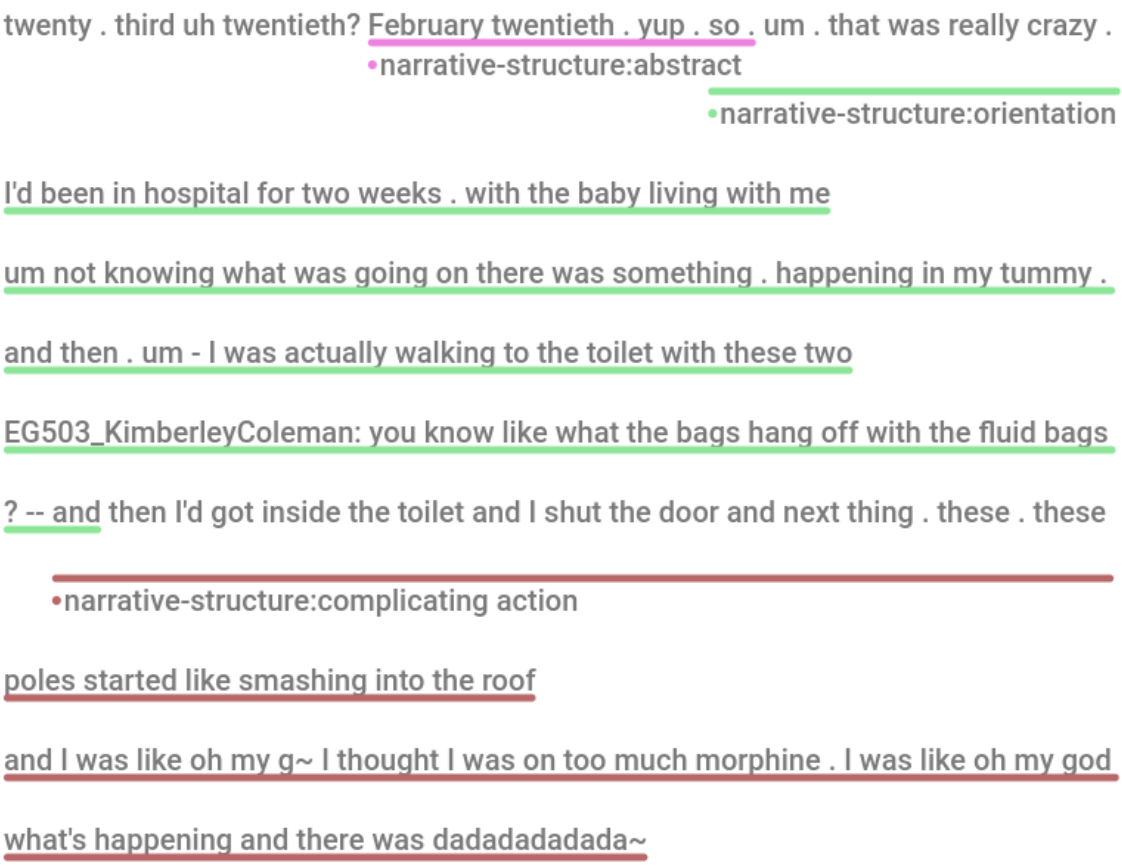


Figure 2: Transcript including annotations exported from LaBB-CAT

The participant ID of the speaker appears at the beginning of each speaker turn, and if you exported phrase/span annotations from LaBB-CAT, they will appear tagging the corresponding regions of the text.

To tag a phrase in the text, simply click and drag over the phrase to select it. A menu of tags will appear.

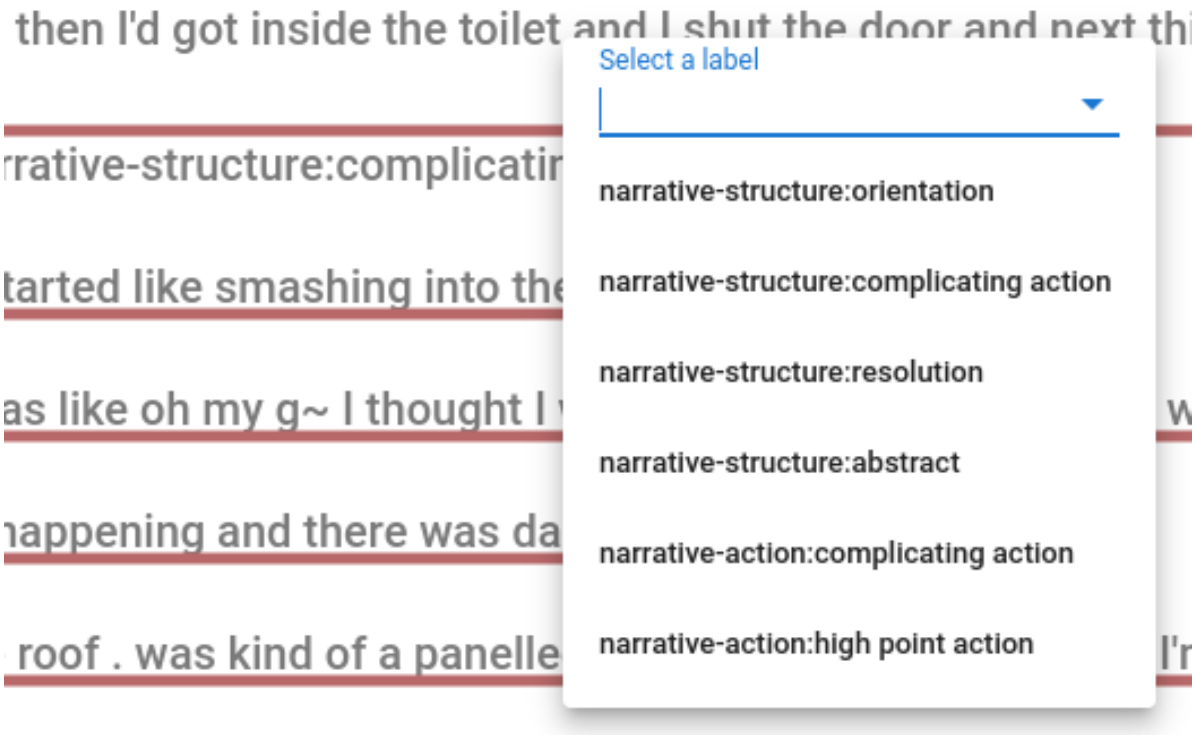


Figure 3: Click/drag for label menu

When you click the desired tag, it will be added to the text.

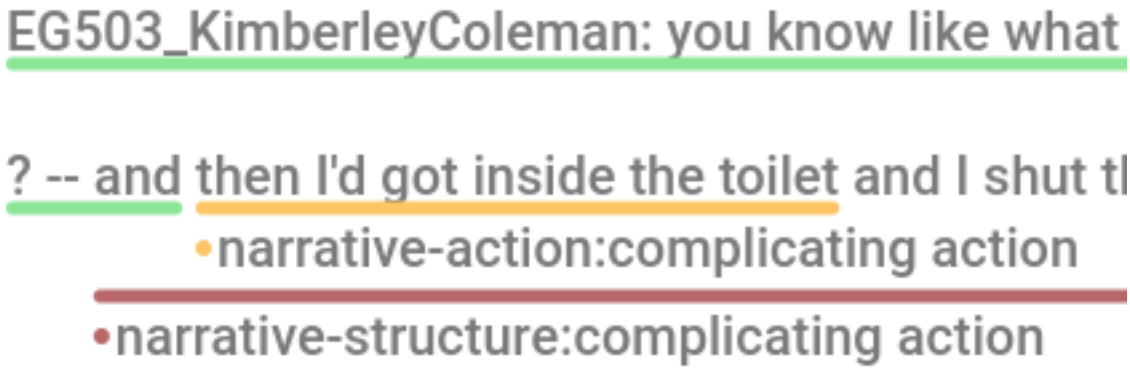


Figure 4: New tags appear in the text

Preselect Label for Tagging

Doccano includes a mode for tagging in which you can pre-select the *Label* you want to use, and then the selected *Label* is automatically used whenever you click/drag a phrase. This mode may be quicker as it involves fewer clicks overall.

To use this method of tagging, scroll to the top of the text, and click the desired *Label* in the list on the top right.

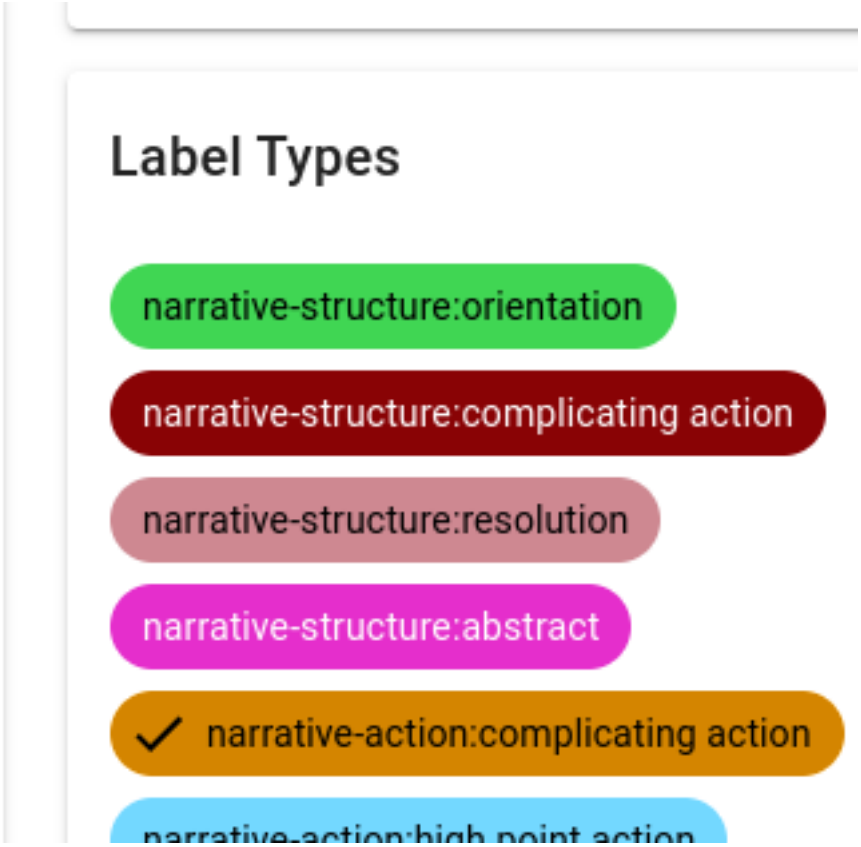


Figure 5: Pre-select a label in the label Types list at the top right of the text

Now, whenever you click/drag a phrase in the tex, it will immediately be tagged with the selected *Label*.

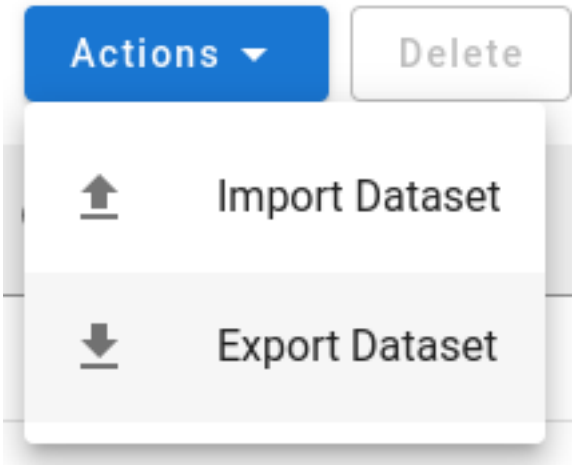
Changes are automatically saved. Once you've added all the tags you want in this text, you can move to the next by using the navigation buttons at the top right of the text.

Figure 6: Buttons for navigating to the first, previous, next, and last text

4. Export from Doccano

One you've finished annotating all texts, you need to export them with the new tags so they can be imported into LaBB-CAT.

1. In Doccano, click the *Dataset* option on the left-hand menu.
2. Move the mouse over *Actions* at the top and select *Export Dataset*.



3. Select *JSONL* as the File format
4. Press *Export*.
5. Save the resulting *…zip* file.
6. Extract the *…jsonl* file that is contained in the *…zip* file you just saved.

5. Import into LaBB-CAT

Ensure phrase/span layers exist in LaBB-CAT

When you import the *…jsonl* file into LaBB-CAT, it will extract the new Labels you've added, and assume that each Label is in the format:
`{LaBB-CAT-Layer-ID}:{LaBB-CAT-Label}`

Each label will be split on the colon, and the left part will be assumed to be a layer ID, and the right part will be assumed to be the label for annotations on that layer.

If you have added layer ID prefixes for layers that don't exist yet in LaBB-CAT, you have to create the LaBB-CAT layers before importing the *…jsonl* file, so that the new annotations have somewhere to go.

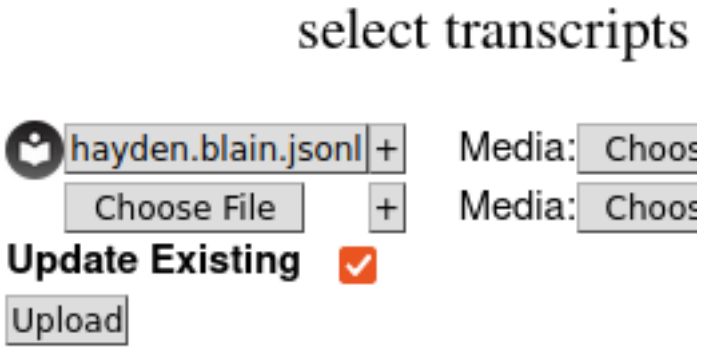
If the new annotations always tag phrases within the same speaker turn (i.e. never cross turn boundaries), then you can add a phrase layer. Otherwise, you must add a span layer.

1. In LaBB-CAT, select *phrase layers* or *span layers* from the menu as appropriate.
2. At the top of the list of layers, fill in the details of the blank row for the layer to add:
    - *Layer ID*: the Doccanno Label's prefix (i.e. the part before the colon)
    - *Type*: *Text*
    - *Alignment*: *Intervals*
    - *Manager*: no manager should be selected
    - *Generate*: Never
    - *Project*: select a project if desired, or none if not
    - *Description*: An informative description of the layer, perhaps including a lilst of all labels included.
3. Click *New* to add the layer

If you have included *Labels* corresponding to multiple LaBB-CAT layers, ensure all the layers have been created in LaBB-CAT before continuing with the import.

Import Dataset

1. In LaBB-CAT, select the *upload* menu option at the top and then the *upload transcripts* option.
2. Press the first *Choose File* button on the left.
3. Select the *…json* file you extracted from the *…zip* file above.
4. Tick the *Update Existing* checkbox.



5. Press *Upload*.
   You will see a list of all the new *Label* prefixes, with a dropdown box for each for selecting the LaBB-CAT layer that the annotations should be imported into.

6. Ensure all *Label* prefixes are matched to the correct LaBB-CAT layer
7. Pres *Next*

Your new annotations will be merged into the existing transcript in LaBB-CAT.

You can double check this by opening on of the transcripts you tagged in LaBB-CAT and ticking the layer(s) of the new annotations. You annotations will appear, lined up with the phrases as you specified in Doccano.