6. Automatic Annotation

You can configure LaBB-CAT to automatically generate annotations, using 'layer managers'. Basically, layer managers are automatic annotation modules that take data in one annotation layer, do some kind of computation on it, and save the result to another annotation layer.

In this exercise, we will use the following layer managers:

- *Frequency Layer Manager*, which counts tokens of each word type, over a configurable scope.
- *Porter Stemmer*, which applies the Porter algorithm to word orthographies to compute word stems.
- *Pattern Matcher*, which creates annotations based on matching regular expressions against words.
- *Statistics Layer Manager*, which computes aggregated information, like word count or duration, over groups of words.

LaBB-CAT comes with number of layer managers pre-installed; you can see a list of installed layer managers by selecting the *layer managers* menu option. Other layer managers have to be manually installed.

For this exercise, we'll pretend we've got a couple more mini-research projects:

- we're interested in looking at how rare or common words are in our data, and
- we want to study 'filled pauses' like "um", "ah", etc.

Frequency

To start with, we'll simply annotate each word token in the database with the count of how many times that word appears in the database...

(1) First of all, create a new project called *frequency*, using the steps we saw before.
(2) Select the *word layers* menu option.
You will see a list of word layers (including the 'custom' layer for the "the" project we created earlier).
(3) Add a new layer, with the following settings:

- **Layer ID:** `frequency`
- **Type:** *Number*
- **Alignment:** *None*

- **Manager:** *Frequency Layer Manager*
- **Generate:** *Always*
- **Project:** *frequency*
- **Description:** `Count of tokens of the same type within each corpus`
  Press the *New* button

(4) You will see the layer configuration form. Fill it in with the following details:

- **Summary:** *Raw Count*
- **Layer to summarize:** *orthography*
- **Scope of Summary:** *Corpus* (leave the box next to that with the *[each corpus]* option selected)
- **Main participants only:** ticked
- **Participants:** un-ticked
- **Filter Layer:** un-ticked
- **Word pairs:** un-ticked
- **Pause Markers:** [leave this blank]
- **Transcript types:** un-tick *wordlist* (as counting word list tokens would artificially inflate frequencies of those words)
- **Annotate tokens:** ticked

> **Note**
>
> If you want more information about what these options mean, check the online help page.

(5) Press *Save*
You will see a message asking you if you want generate the layer data now.
(6) Press *Regenerate*.
You will see a progress bar moving across the page while the counts are being generated. When it is finished, you will see a message saying "Layer complete…"

Now each word in each transcript is annotated with the count of the number of instances of that word with the corpus of the transcript.

To see what that looks like…

(7) Select the *transcripts* menu option.
(8) Select the name of the first transcript in the list.
(9) At the top of the transcript there is now a list of projects. Tick the "frequency" project.
This will reveal the frequency layer in the list of layers.
(10) Tick the *frequency* layer.
When the transcript reloads, you will see that above each word is a number. That number is the number of times that word appears in the transcript's corpus.

e.g. if the word “and” has 669 above it, and the transcript is in the *QB* corpus, that means that the word “and” appears in the *QB* corpus 669 times.

The newly-generated annotations are also searchable…

(11) Select the *search* menu option.
(12) If the *frequency* project and layer are not already ticked, tick them to add the *frequency* layer to the search matrix.

In the search matrix, you will notice that, unlike the *orthography* layer, which has one box for a regular expression, the *frequency* layer has two boxes, marked “≥” and “<”. For a layer of type *Number* (which is what you specified above), instead of a regular expression, you can match by numeric range.
(13) We want all the words that appeared only once in their corpus. Enter a number or numbers in the appropriate box (you can leave either box blank) and press *Search*.
(14) Press ◪ *20 More Matches* a couple of times, to get a good idea of the range of results.

The results you see may contain words that don’t seem rare at all. That they only appear once is a product of two factors:

i) there isn’t that much data in our example database, and

ii) these are counts of ‘wordforms’ - i.e. the surface spelling of the word; e.g. the word “damaging” might be quite rare, even though there are more instances of words from the same stem like “damage”, and “damaged”. This second factor will be addressed soon…

You can also extract the annotations into CSV results from other searches…

(15) On the search page, do a search for “the” followed by a word that starts with a vowel.
(16) When the results page appears, press the ◪ button next to the *CSV Export* button.
(17) Under the list of *Word* layers, tick the *frequency* layer.
(18) Press the *CSV Export* button.
(19) Save and open the resulting CSV file.
You will notice that in the spreadsheet there are two columns:

- *Match frequency:* this lists the frequency of each word that matched, in order. i.e. in this case two numbers, the frequency of “the”, followed by the frequency of the word after it.
- *Target frequency:* this contains a single frequency, in this case the frequency of the first word that matched a pattern - i.e. “the”

(20) As an aside, you can also select other layers to include in the CSV file. For example, some of the transcripts include topic-tags that were made in the original ELAN transcript.

Export your search results to CSV again, this time including the *topic* layer, and see what that looks like.

The Frequency Layer Manager also keeps a word-list with token counts for each corpus…

(21) Select the *layer managers* menu option.
(22) On the "Frequency Layer Manager" row, press the *Extensions* button.
(23) You will see a drop-down box with each corpus in it.
Select *QB* and press *Export*.
(24) Save and open the resulting CSV file.
You will see an alphabetical list of all the distinct word types in the QB corpus, and next to each, a count of the number of tokens of that type in the QB corpus.

Porter Stemmer

As pointed out above, although the 'wordform' counts might be useful, it also may be useful to lump together different forms of the same stem for the counts. e.g. if there's 1 "damaging" token, 28 "damage" token, and 18 "damaged" token, it may be useful to count these all together as 47 tokens of the same stem.

In order to achieve this, we first need to 'stem' all the words in the database - i.e. reduce all the wordforms so that tokens like "damaging", "damage", "damaged", and "damages" all have the same 'stem' annotation. Then we can gather frequency statistics on the stems.

The *Porter Stemmer Layer Manager* is one way to achieve this. First, we need to install this layer manager (which only works on English data, so it's not installed by default).

(25) Select the *layer managers* menu option.
(26) Near the bottom of the page, select the *List of layer managers that are not yet installed* link.
(27) Find the "Porter Stemmer" in the list, and press its *Install* button, and then *Install* again to continue.
(28) After it is installed, a tab appears with some information about what the layer manager does. You may wish to read this page for your information. Afterwards, you can close the tab to take you back to the LaBB-CAT browser tab.
(29) Select the *word layers* menu option.
(30) Add a new layer with the following attributes:

- **Layer ID:** `stem`
- **Type:** *Text*
- **Alignment:** *None*
- **Manager:** *Porter Stemmer*
- **Generate:** *Always*

- **Project:** *frequency*
- **Description:** `The stem of the word according to the Porter algorithm`

  Press the *New* button.

(31) The Porter Stemmer's default configuration is fine for our purposes, so press *Set Parameters*.
(32) Press *Regenerate*.
  You will see a progress bar, and once it's finished, you will see a message saying "Layer complete…"
(33) Select the *transcripts* menu option.
(34) Click the name of the first transcript.
(35) Tick the *stem* layer we just added (you may need to tick the *frequency* project to reveal the *stem* layer).
  When the transcript refreshes, you will see, above each word, its 'stem' according to the Porter algorithm.

You will notice that, although the stems are not what you might regard as being the 'lemma' of each word (i.e. not necessarily valid words of English in themselves), they nevertheless generally strip off plural and 3rd-person-present suffixes, such that different wordforms of the same lemma will have the same 'stem'.

Now that we have generated a layer of 'stems' for the wordforms on the *orthography* layer, we can generate frequency data from the *stem* layer as well…

(36) Click the *word layers* menu option.
(37) Add a new layer with the following attributes:

- **Layer ID:** `stemFrequency`
- **Type:** *Number*
- **Alignment:** *None*
- **Manager:** *Frequency Layer Manager*
- **Project:** *frequency*
- **Description:** `Count of tokens of the same stem within each corpus`

  Press the *New* button.

(38) Configure the layer exactly as before, except this time, set the **Layer to Summarize** setting to the *stem* layer we created above. *Save* your settings and press *Regenerate*.
  The layer will be generated.
(39) Do a search of all speakers, for words with a value of 1 on your *frequency* layer, but a value of more than 1 on your new *stemFrequency* layer.
  You will see all the words that were previously shown as 'rare', but are actually more common if you count all the tokens with the same stem together.

Pattern Matcher

We will now create some automatic annotations of a different kind. Let's suppose that we're interested in 'filled pauses' – words like "um", "ah", "er", "mmm", etc. You can actually identify them using regular expressions…

(40) Do a search of all speakers, for the word `ah`. Select the *no matches, only a summary of results* option.
Note the number of results you get back.

(41) Now do a similar search, for the pattern: `a+h+` i.e. `1` or more *a*'s followed by one or more *h*'s.
Note the number of result you get back is more than in the previous search. It turns out the transcribers, when transcribing the word "ah" weren't entirely consistent in their spelling of that word. That's ok, because with a little imagination, we can invent searches that will identify filled pauses like "um", "ah", and "mm", even if they've been spelt "umm", "ahh", or "mmm".
(It turns out that there's a good reason to prefer "mmm" over "mm", but we'll see that in a later exercise)

(42) Try out a few different searches to see if you can identify different ways that transcribers have spelt filled pauses like this.

We could annotate these as filled pauses by searching, annotating a CSV file, and uploading the CSV annotations, as we did previously. However, there is a layer manager that can do this for us, for all the existing data, and for any new transcripts that might be uploaded in the future: the "Pattern Matcher" layer manager.

(43) First of all, create a new project called *pauses*.

(44) Now create a new *word layer*, with the following attributes:

- **Layer ID:** `pause`
- **Type:** *Text*
- **Alignment:** *None*
- **Manager:** *Pattern Matcher*
- **Generate:** *Always*
- **Project:** *pauses*
- **Description:** `Filled pauses annotated by regular expression`

Press the *New* button

(45) Set the **Source Layer** to be *orthography*.
The **Destination Layer** and language-related settings can be left with their default values.

Below this, there is a currently empty list of "Mappings". We are going to add regular expressions to this list, which will identify filled pauses.

(46) On the new empty row that's already in the list by default, select the box labelled "Source pattern", and enter: `u+m+`
(47) To the right of this, select the "Destination Label" box and enter: *um*

This will make the layer manager find any instances of words that match the pattern "*u+m+*" on the *orthography* layer, and in each case, save the annotation"um" on our new *pause* layer.

(48) Press the + button to add a new blank row, and add another regular expression:

- **Source Pattern:** `a+h+`
- **Destination Label:** `ah`

(49) Press the + button again, and add another regular expression:

- **Source Pattern:** `mm+`
- **Destination Label:** `mm`

(50) Add any more regular expressions you think might help identify filled pauses.
(51) Under the patterns, select the option to *Delete annotations in target layer whose source matches no pattern*

> **Tip**
>
> If you would like more information about the pattern configuration and what kinds of target annotations you can create, you will find that clicking on the brief description of the layer manager above the form expands to provide more detail.

(52) Press *Set Parameters* and *Regenerate* to generate the layer.
You will see a progress bar while the layer manager annotates all the filled pauses in the database.
(53) To see what this looks like in a transcript, perform a search for *um* on your new *pause* layer, and click on the first match.
You should see that each instance of the word "um" (or its variants) has been annotated, as have instances of "ah" and "mm".

Now that these filled pauses are automatically annotated, there are various things you might do with the annotations. You could:

- include them in the context of multi-word searches, for example you might want to study the effects of a filled pause on the following or preceding word, or
- search for only the pauses themselves, for selected speakers, in order to study what kinds of filled pauses are used by which speakers in what contexts, what their durations are, etc.

Statistics Layer Manager

In fact, we can use another layer manager to automatically count them for each speaker, and for each utterance in the transcript. In order to do this, we are going to create a 'phrase layer', which is a layer that can contain annotations over groups of words (as opposed to against individual words). The layer manager we will use can also annotate participants…

(54) Select the *phrase layers* option on the menu.
You will see a list of phrase layers that are already set up, including *language* and (named) *entity*.

(55) Add a new layer with the following characteristics:

- **Layer ID:** `pauseCount`
- **Type:** *Number*
- **Alignment:** *Intervals*
- **Manager:** *Statistics Layer Manager*
- **Generate:** *Always*
- **Project:** *pauses*
- **Description:** `Count of filled pauses, for the utterance and the speaker`

Press *New*

(56) You will see a form for the layer's configuration. Fill in the details as follows:

- **Layer to summarize:** *pause*
- **Statistic:** *Token Count*
- **Pattern to match:** [leave this blank]
- **Context:** [leave this blank]
- **Pause Threshold:** [leave this blank]
- **Main-participant utterances only:** ticked
- **Scopes:** tick *Utterances*, and under *Participants:*, select the option *add new attribute called pauseCount*
- **Transcript types:** leave all the options ticked

> Tip
>
> If you would like more information about what these settings and the other options do, try the online help for this page.

(57) *Save* the layer configuration, and then press *Regenerate*.
You will see a progress bar while the layer manager annotates all the transcripts in the database.

(58) To see what this looks like in the transcripts, select the *transcripts* option on the menu, clear all filters, and open the first transcript in the list.

(59) Under the list of projects, if the *pauses* project isn't already ticked, tick it, which will reveal the *pauseCount* layer in the list of layers.
(60) Tick the *pauseCount* layer.
Scrolling down the transcript, you will see that, wherever there is a filled pause like "um", the entire utterance in which it appears has a bracket across the top of the words, labelled with the number of filled-pauses that occurs in that utterance.
(61) Scroll to the top of the transcript, and click on the name of the main participant.
You will see the participant's attributes page, which now includes the participant's *pauseCount* attribute.
(62) Both the local utterance count, and the participant's overall count, can also be exported to CSV search results files.
Select *search* and perform a search involving the *pause* layer.
(63) At the bottom of the results page, press the ◪ button next to the *CSV Export* button, to reveal the layer options.
(64) Under *Participant* layers tick the *pauseCount* attribute.
(65) Under *Phrase* layers tick the *pauseCount* layer.
(66) Press *CSV Export*, and save and open the resulting file.
You will notice that there is a column called "participant_pauseCount" with the participant's global count, and another called "Target pauseCount" with the local utterance count.

The Statistics Layer Manager can also incorporate time information in its computation, so it can be used to compute speech-rate. We could use it on our example database to compute words-per-minute for utterances, turns, speakers, etc.

If you like, you can try to figure out how to set up a "words-per-minute" layer now.

However, normally speech-rate is expressed in syllables per minute. We don't have any way to get syllable-counts for our words yet, but we will be doing that in a later exercise...

In this exercise, you've seen how layer managers can be used to compute new annotations automatically from existing annotations, e.g.

- Words can be tagged with their frequency in the LaBB-CAT database, or its corpora.
- Words can be tagged with their 'stem' using the Porter Stemmer.
- Words can be tagged with annotations on the basis of regular expressions.
- Groups of words can be tagged with aggregated information like word count or rate over time.