Part of Speech Tags

LaBB-CAT can be integrated with the Stanford POS Tagger, which is free software developed by The Stanford Natural Languages Processing Group for tagging words in various languages with their parts of speech.

In this exercise you will:

1. install the Stanford POS Tagger layer manager module, and
2. use it to tag each word with its part of speech.

Install the Stanford POS Tagger

The first thing we're going to do is install the Stanford POS Tagger layer manager, which is a LaBB-CAT module that integrates with the Stanford NLP Group's software…

(1) In LaBB-CAT, select the *layer managers* menu option.
You will see a list of pre-installed layer managers, which are modules that can perform automatic annotation tasks. The Stanford POS Tagger layer manager isn't pre-installed, because it is language-specific, and requires installation of further software.
(2) Near the bottom of the page there a link labelled:
List of layer managers that are not yet installed – click it.
(3) Find *Stanford POS Tagger* in the list, and press its *Install* button.
(4) Press *Install* on the resulting information page.
This displays some further information about the layer manager, allowing you to optionally upload an alternative version of Stanford's software.
We won't upload a file, we'll be using the standard file that is included in the layer manager.
(5) Press *Configure*.
You will see a progress bar while the layer manager downloads the software from the Stanford website. This will take a minute or so.
(6) Once it's finished, you will see a new window open with information about the Stanford POS Tagger layer manager.

Annotate Words with Part of Speech tags

Now that we've installed the layer manager, we'll create an annotation layer that tags words with their pronunciations.

(7) Select the *word layers* option on the menu.
You will see a list of existing word layers, including the *orthography* layer, the *lexical* layer, etc.

(8) The column headings are also a form for defining a new word layer. Fill in the following details in this form:

- **Layer ID:** `pos`
- **Type:** *Text*
- **Alignment:** *Intervals*
- **Manager:** *Stanford POS Tagger*
- **Description:** `Part of Speech tag(s) according to the Stanford POS Tagger.`

(9) Press *New* to add the layer.
You will see the layer configuration form.

There are some word tokens we want the POS tagger to ignore:

- filled pauses like "um", "ah", and "mm", and
- half-finished words that the speaker interrupted before completing the full word - these are transcribed with a ~ at the end of the word, e.g. if the speaker started saying "noise" but changed their mind before the end of the word, this might be transcribed as "noi~".

This is what the *Token Exclusion Pattern* setting is for; it's a regular expression that identified words that should be excluded from part-of-speech tagging.

(10) Set the *Token Exclusion Pattern* to be:
`um|ah|mm|.*~`

> **Tip**
>
> If you're curious about what the configuration options do, hover your mouse over each one to see further information about what the setting does.

(11) Press *Set Parameters*.
You will see a message asking you if you want (re)generate the layer data now.
(12) Press *Regenerate*.
You will see a progress bar moving across the page while the annotations are being generated. This will probably take a minute or so.
When it is finished, you will see a message saying *Finished*.
(13) Select the *transcripts* menu option, and open the first transcript in the list by clicking the transcript name.
(14) Select the *Layers* tab at the top to reveal a list of tickable annotation layers.
(15) Tick your new *pos* layer.

You'll see that each word is now tagged with at least on part-of-speech tag. Some words will have multiple tags, for example "I've" includes

- a `PRP` (personal pronoun) and
- a `VBP` (present-tense verb).

These tags (like any annotations in LaBB-CAT) can be searched, extracted, and analysed.