Stanford POS Tagger

Depending on the language of your transcripts, you may be able to tag each word with its part of speech (Noun, Verb, Adjective, etc.) using the Stanford POS Tagger.

The Stanford POS Tagger has models for:

- Arabic
- Chinese
- English
- French
- German
- Spanish

The steps for POS tagging your corpus are:

1. Install the Layer Manager
2. Configure a POS layer

Install the Layer Manager

1. In LaBB-CAT, select the *layer managers* option on the menu at the top.
2. At the bottom, follow the link labelled: *List of layer managers that are not yet installed.*
3. Find the *StanfordPosTagger* layer manager in the list, and press its *Install* button, then *Install* again.
   You will see a configuration page with some information about the tagger.
4. Press *Configure.*
   You will see a progress bar while the layer manager downloads the Stanford POS Tagger files.

Once it's finished, you'll see a further information page.

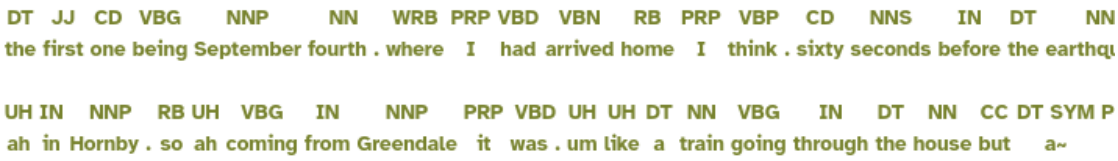
Create a POS layer

Now the layer manager is installed, we need to create a layer that is configured to use it to tag words with their part of speech…

1. Select *word layers* on the menu at the top.
2. You will see a list of word tag layers that have already been configured. The column headings at the top are also a form for creating a new layer, so we'll fill in that form now.
3. Fill in the following details on the form at the top:

- *Layer ID*: `pos`
- *Type*: *Text*
- *Alignment*: *Intervals*
  NB it's important that this is not set to *None* because a single word can have multiple POS tags, one after another, which are strung between the start and the end of the word token.
- *Manager*: *Stanford POS Tagger*
- *Generate*: *Always*
- *Project*: This can be left as the default value, unless you want to add the layer to a category of your choice.

4. Press *New*
   You will see the layer configuration form. Mostly you can leave the default values as they are.
5. Set the *Model to use* setting to something that makes sense for your transcripts, which depends on their language. This is a setting you may experiment with to get the best results. For English recordings, you may find the *english-bidirectional-distsim.tagger* is slower but produces better results.
6. Press *Set Parameters*.
7. Now press *Regenerate* to run the POS tagger on your whole corpus.
   You will see a progress bar while the transcripts are being tagged.
8. Once it's complete, select the *transcripts* option on the menu, and click the first transcript in the list.
9. Tick the new *pos* layer to display the tags.
   You will see that each word has one or more tags above it - these identify the parts of speech or syntactic categories of the words.



The tags can be searched, extracted, summarised, etc. just like any other annotations.

Summarising POS Tags

One possible aggregated analysis might be to compute the distributions of POS tags for each speaker.

In order to do that, you would set up the tagger as above to output the POS tags, and then set up a Frequency Layer Manager layer with the POS tags as input. In order to do that:

1. In LaBB-CAT, select the *word layers* menu option.

2. Add a new word layer with the following characteristics:

    - *Layer ID*: `posFrequency`
    - *Type*: *Number*
    - *Alignment*: *None*
    - *Manager*: *Frequency Layer Manager*
    - *Generate*: *Always*

3. Press *New*.
   You will see the Frequency Layer Manager configuration form.
4. Configure the layer as follows:

    - *Summary*: *Raw Count*
    - *Layer to Summarize*: *pos* (i.e. the POS layer we created earlier)
    - *Scope of Summary*: *Speaker*
      The rest of the settings can be left with their defaults, except:
    - *Annotate Tokens*: unticked - we only want the summary information.

5. Press *Save*.
6. Press *Regenerate* to analyse your corpus.
   You'll see a progress bar while each POS label is counted for each participant.
7. Once the layer manager is finished, select *layer managers* on the menu.
8. Find the *Frequency Layer Manager* in the list, and press its *Extensions* button.
   This will show a page that lets you select from 'dictionaries' that are named after layers managed by the Frequency Layer Manager.
9. Select *posFrequency* and press *Select*
   A form is displayed that allows you to perform various operations on the frequency lists you have generated. Most likely, you just want to export a list of frequencies for a speaker:
10. Under *Scope*, select a speaker.
    Or if you want to include all speakers, select the *[all scopes]* option at the bottom of the dropdown box options.
11. Press the *Export* button at the bottom.
    This will give you a CSV file. If you open this in Excel (or any other data analysis tool), you'll see that it contains three columns:

    - *Scope* - the speaker name
    - *Type* - the POS label
    - *Frequency* - the number of times that speaker uttered a word with that POS label

1	Scope	Type	Frequency
2	AP511_MikeThorpe	CC	71
3	AP511_MikeThorpe	CD	10
4	AP511_MikeThorpe	DT	96
5	AP511_MikeThorpe	EX	8
6	AP511_MikeThorpe	IN	125
7	AP511_MikeThorpe	JJ	54
8	AP511_MikeThorpe	JJR	3
9	AP511_MikeThorpe	JJS	5
10	AP511_MikeThorpe	MD	12
11	AP511_MikeThorpe	NN	89
12	AP511_MikeThorpe	NNP	57
13	AP511_MikeThorpe	NNPS	2
14	AP511_MikeThorpe	NNS	30
15	AP511_MikeThorpe	PDT	1
16	AP511_MikeThorpe	POS	2
17	AP511_MikeThorpe	PRP	107
18	AP511_MikeThorpe	PRP$	14
19	AP511_MikeThorpe	RB	121
20	AP511_MikeThorpe	RBR	2
21	AP511_MikeThorpe	RBS	1
22	AP511_MikeThorpe	RP	8
23	AP511_MikeThorpe	SYM	10
24	AP511_MikeThorpe	TO	16
25	AP511_MikeThorpe	UH	57
26	AP511_MikeThorpe	VB	29
27	AP511_MikeThorpe	VBD	108
28	AP511_MikeThorpe	VBG	23
29	AP511_MikeThorpe	VBN	29
30	AP511_MikeThorpe	VBP	22
31	AP511_MikeThorpe	VBZ	10
32	AP511_MikeThorpe	WDT	11
33	AP511_MikeThorpe	WP	6
34	AP511_MikeThorpe	WRB	12
35	AP513_Steve	CC	35
36	AP513_Steve	CD	9

If you prefer to have POS counts by transcript instead of by speaker, you can select *Transcript* as the scope at step 4 above. If you want both of these, create two word layers, one for summarising by participant, and the other by transcript.