Exercise 5   Syntax and the Stanford Parser

The Stanford Parser is an open-source PCFG parser that can use grammars for a variety of languages including English.

LaBB-CAT includes a Layer Manager that handles integration with the parser. The Stanford Parser Layer Manager:

- extracts chunks of transcripts (ideally sentences or clauses),

- gives them to the Stanford Parser for processing, which produces a 'best parse' for the utterance provided, and

- saves the parse on a 'tree' layer, and optionally saves the resulting part-of-speech tags on a word layer.

One of the problems with parsing speech is that speakers often don't speak in complete, well-formed sentences. In addition, the demo corpus you are using was not generally transcribed with parsing in mind, and so grammatically complete units have not been marked with full-stops, commas, etc. (Instead, full-stop has been used to mark short pauses in speech).

For these reasons, the parses you will see in this data may not be perfect. However, it's possible to get a sense of the kinds of things that could be achieved with well-formed written texts, or speech that has been transcribed with grammatical punctuation included.

1. Click the *transcripts* link on the menu.
   One transcript in the database has delimiters inserted which divide the transcript into more or less grammatical units. As the 'full-stop' symbol is already being used to mark pauses, the 'vertical bar' symbol | has been used as a grammatical delimiter.
   The transcript is called "BR2044_OllyOhlson.eaf"

2. In the *Transcript:* box, type `olly` and hit < Enter >.
   "BR2044_OllyOhlson.eaf" will be the only transcript in the filtered list.

3. Click *BR2044_OllyOhlson.eaf*

4. Untick all except the *transcript* layer, to avoid clutter.

5. When the transcript re-loads, tick the 'syntax' project.
   This reveals three layers.

6. Tick the *parseable* layer.

When the transcript re-loads, you will see that almost all words in the transcript have been tagged with their own orthography. However, some words have not been tagged:

- filled pauses like "um", "ah", etc. and

- incomplete words like "re~", "na~", etc. − i.e. cases where the participant started saying something but changed their mind.

These have been identified by the Pattern Matcher Layer Manager, which can pick out words by regular expression, and has been configured to tag as 'parseable' all tokens *except* those that matching the following patterns:

- `a+h+` − e.g. "ah", "aah", "ahh", . . .

- `m+h*m+` − e.g. "mm", "mmm", "mhmm", . . .

- `e+r+` − e.g. "er", "err", "eeerr", . . .

- `u+m+` − e.g. "um", "uum", "ummm", . . .

- `.+~` − e.g. "re~", "na~", "w~", . . .

The result is that the *parseable* layer includes all words except filled pauses and incomplete words. This is the layer that is passed to the Stanford Parser for syntactic parsing.

One of the results of parsing is that each word token is tagged with its part-of-speech.

7. Untick the *parseable* layer and tick the *pos* layer.
   You will see that most of the words have been tagged with a syntactic category:

- "CC" Coordinating conjunction
- "DT" Determiner
- "JJ" Adjective
- "NN" Noun
- "NNS" Plural noun
- "NNP" Proper noun
- "PRP" Personal pronoun
- "PRP$" Possessive pronoun
- "VB" Verb
- "VBD" Past tense verb
- "VBG" Gerund
- "VBZ" 3rd person singular present verb
- . . . etc.

These, like any other word tags you have seen, can be included in searches or exported to CSV results files.

You may notice that, in addition to the filled pauses and interrupted words, there are some other lines that have no part-of-speech tags. These are utterances in Te reo Māori; these utterances have been manually marked as in a different language – if you tick the *language* layer, you can see the language annotations. The Stanford Parser Manager has been configured to parse only English utterances, so the Te reo Māori ones are skipped.

The other result from parsing is, of course, a 'parse tree' of each utterance.

8. Untick the *pos* layer (and the *language* layer if it's ticked) and tick the *parse* layer.
   You will see that above the words, there are bracketing annotations that are labelled with parts-of-speech or phrase labels.
   Each of these brackets represents a syntactic constituent constructed by the parser, smaller constituents at the bottom, building into larger constituents going up.

9. Click on any constituent label (e.g. "NP" or "S").
   A new window will open, which shows the selected utterance, using the familiar 'upside-down tree' representation.
   If the tree appears far off and small, you can make it larger by widening the window, or 'zooming in' with you mouse wheel.

You can also search the parses themselves.

10. Close the parse tree window.

11. Click the *search* option on the menu.

12. In the *Series:* dropdown box, select the 'BR2044_OllyOhlson' option.

13. When the page reloads, press *Layered Search.*

14. Tick the 'syntax' project.

15. Tick the *parse* and *pos* layers amd press *Set Search Matrix.*

On the *parse* layer you'll see that you can enter a search expression for annotations on that layer, just like any other.
   However, it also has a checkbox before and after the pattern. If you tick the checkbox before the pattern, it will anchor the search to the first word in the matching contituent. Similarly the checkbox after the pattern anchors to the last word in the constituent.
   Let's say you want all the noun phrases that *don't* start with a determiner like "the", "this", "a", etc.

16. Enter a pattern that matches NP (noun phrase) on the *parse* layer. . .

17. . . . and tick the checkbox to anchor to the first word in the consituent.

18. Enter a pattern that would match DT (determiner) on the *pos* layer...

19. ...and in the dropdown box before the pattern, select 'not'.
    This will match words who have an annotation that doesn't match "DT" on the *pos* layer.
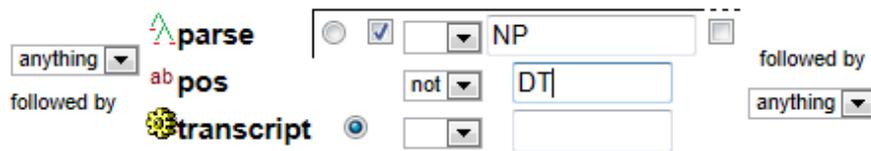
20. Press *Search*.
    When the search finishes, you should see that lots of nouns are returned, but nothing preceded by "the", "a", "that", or any other determiner.

In this worksheet you have seen that:

- the Stanford Parser can we used to annotate transcripts with part-of-speech tags and constituent annotations, and

- the resulting annotations can be included in syntax-based searches.