Exercise 6 Other Processing

1 Aggregate Measures

You have seen in a previous worksheet that articulation rate can be calculated over the words in individual utterances, and also over all the words uttered by each participant. There are other useful computations that can be computed over different scopes.

1. Click the *transcripts* link on the menu.
   You may have previously noticed that the top of the page includes a form with various transcript attributes. This form allows you to both filter and sort the list of transcripts by transcript attribute values.

2. For the *Word Count:* attribute, there are two boxes which you can use to specify a range of values.
   In the right-hand *To* box, enter `1000` ~~and hit < `Enter` >~~
   You will see a list of all transcripts that have up to 1000 words.
   This word count was computed by the Statistics Layer Manager, which has also been configured to compute speech duration in seconds and save the result in the "Duration" transcript attribute.

3. ~~Click the "Duration" attribute (i.e. the label that says "Duration")~~
   ~~You will see that the durations of the transcripts are displayed in a column, and that the transcripts are ordered by duration.~~

4. ~~Click the "Duration" attribute again.~~
   ~~You will see that the order of the list is reversed; instead of being shortes to longest, it's now longest to shortest.~~
   This functionality doesn't seem to be there anymore.

Another simply aggregate calculation is type/token ratio. The Demo LaBB-CAT has been configured to compute the type/token ratio for each participant.

1. Click the *search* link on the menu.

2. For the *corpus:* attribute, select the 'QB' option to list only participants recorded in the "Quake Box" portable recording studio.

3. At the bottom of the participant list, press the *Export* button.

4. Tick the 'Gender', 'Age' and 'type/token ratio' attributes.

5. Press the *Participant Data* button.

6. Save and open the resulting CSV file.
   You will see that you have a list of their participants, with gender and age, and also a column for type/token ratio; this is the ratio expressed as a percentage.

2 Other Media

The transcripts in this database each have a video and an audio file.

However, some of the recordings have also been processed with a facial feature location algorithm. One of the results of this process was an annotated video; a copy of the original video, with the participant's face located, along with various facial landmarks (position of the eyes, shape of mouth, etc.).

LaBB-CAT supports having multiple media 'track' files for the same transcript, and for some of the transcripts, the annotated video has been uploaded as well as the original video.

7. On the *transcripts* page, list the transcripts with the *Quake Face:* attribute set to 'true'.

8. Open one of the listed transcripts.
   By default, the original video is selected for display, but all the other media files available for the transcript are listed below the video, with a checkbox next to each.

9. Tick the checkbox next to the media file that ends with "..._face"
   The transcript will reload and display the annotated video.

10. Press *Play* to see the annotations (marked in green and red) change with the facial features.
    It may be easier to see the annotations if you put the video in 'full screen' mode.

3 Keyness

In addition to calculating word frequencies for direct analysis, frequencies can be compared to a reference corpus to calculated their ‘keyness’; a measure of whether the word is unusually frequent (a high positive keyness) or unusually infrequent (a low negative keyness).

The Demo LaBB-CAT has been configured to compute keyness compared to the frequencies available in the CELEX lexicon, which come from the Cobuild corpus.

11. Click the *home* link on the menu.
12. Click the ‘Keyness’ icon.
    You will see a form that allows you to search for particular spelling patterns, or export a list.
13. Press the *Search* button without filling in the *Pattern:* box, to list all words above the default *Keyness:* threshold.
    A list of words will be displayed, each word with its keyness metric. The high-positive words (which are unusually frequent) are listed first, with the low-negative words (unusually infrequent) below.

Unsurprisingly for this speech corpus, as compared to the mostly-written Cobuild corpus, words with high keyness include filled pauses like “um” and “ahh”, other words more likely in informal speech like “gonna” and “yeah”, topic-specific words like “earthquake” and “aftershocks”, and Canterbury place-names like “Christchurch” and “Brooklands”.

The Frequency Layer Manager can be configured to compute keyness of the data compared to any corpus for which you have word frequency data, or if you have several corpora within one LaBB-CAT database, each corpus can be compared to all the rest.

4 Linguistic Inquiry and Word Count

Linguistic Inquiry and Word Count (LIWC) text analysis can be done with the LIWC Layer Manager and categorised word lists.
See: https://liwc.wpengine.com/how-it-works/
Or: Tausczik & Pennebaker (2010) "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods" Journal of Language and Social Psychology 29 (1) 24-54

LIWC involves calculating the percentage of words in different categories. Categorised word lists can be purchased from liwc.wpengine.com, or can be compiled by hand.

LIWC text analysis has been done on the Demo LaBB-CAT database, and also on the Cobuild corpus as a comparison corpus.

14. Click the *home* link on the menu.
15. Click the ‘LIWC’ icon.
    You will see a horizontal bar graph: each bar represents category of words, with the bar length representing the percentage of that category’s usage in the database.
16. Tick the ‘Cobuild’ checkbox.
    Bars representing the percentages for the Cobuild corpus will be added to the graph, for comparison.
17. Press the *Export* button.
18. Save and open the resulting CSV file.
    You will see that the file contains the list of categories, with two percentages for each category, first the percentage for the LaBB-CAT data, and then the percentage for the Cobuild corpus.

5 Personality

LaBB-CAT can also integrate with the IBM Watson Personality Insights service.
(https://www.ibm.com/watson/services/personality-insights/)
This is a web service that, given texts, provides personality metrics on the author (or speaker) of the text.

The transcripts in the Demo LaBB-CAT have been processed by the Personality Insights service. The results can be listed and visualised per speaker.

19. Click the *home* link on the menu.

This wasn't working when I tried it.

20. Click the ‘Personality’ icon.
    You will see a list of participants that have been analysed.

21. Click on the name of a participant.
    You will see a ‘sunburst’ style visualisation of the participant’s personality metrics.
    Below the visualisation, the categorised metrics are listed in a table.

In this worksheet you have seen that:

- the Statistics Layer Manager can provide a variety of summary computations,
- the transcript list can be filtered and sorted by transcript attributes,
- transcripts can be linked to multiple media files,
- unusually frequent or infrequent words can be identified,
- LIWC text analysis can be automatically performed, and
- personality metrics can be obtained for participants, based on their utterances.