Cleanup: Batchalign output will need to be checked and cleaned up in various ways.

For all clean-up, you should be opening the *.cha output files using the CLAN editor. The "cha" extension stands for CLAN. CLAN can be downloaded from https://dali.talkbank.org. Once CLAN is installed, you can double-click on a *.cha file and it will open in the CLAN editor.

- The output of **transcribe** will need thorough checking as described below.
- The output from **align** should be checked by running CHECK. You can type escape-L in an open file to run CHECK on the file.
- The output from **morphotag** will pass CHECK, but you will want to examine the results for lexical and grammatical accuracy. This type of tagging is never 100% accurate.

** Checking and cleanup after **transcribe** involves playing back the audio as it is linked to the transcript on the utterance level. To do this, the *.cha file and the *.wav file must be in the same folder. You then type escape-8 in CLAN to start this continuous playback procedure. You can stop continuous playback with a click and play individual utterances by command-click on the main line.

** If you ran transcribe using Rev-AI, it will create speaker IDs with names such as PAR0 and PAR1. Before running correction with continuous playback, it is best to use the ROLES command in CLAN to convert these speaker codes to things like INV and PAR, based on checking the a few lines of the file to see how to best assign these. You can consult the CHAT manual section 7.2 for the list of possible roles. Before running ROLES, you should make a backup version of your files. Then, the command is: roles +c *.cha +1 which uses a roles.cut file like this:

```
PAR0  INV          Investigator
PAR1  PAR          Participant
```

** The next steps are done using CLAN's continuous playback command (Esc-8) to play back and listen through the whole transcript. This will take some time. As you go, you will want to fix these things.

1. Utterance termination decisions -- NLP analysis requires that each utterance be on its own line
    - split utterances where needed. One way of doing this is to insert a period after each utterance and then to use CLAN's FIXIT program to break up larger blocks into single utterances.
    - join utterances where needed. This is done by cutting out the speaker ID and carriage return.
2. Utterance termination markers to replace periods: ? ! +...
3. Speaker ID header fixing. Running ROLES may fix some of this, but you can use the Tiers pulldown menu in CLAN to fix individual utterances by typing Command and the number.

Make sure you run "Update ID headers" at the bottom of the menu to get the current assignments.
4. Accuracy of words transcribed. ASR will often produce an incorrect word that is close to the target. You need to replace this with the correct word.
5. Revision and repetition coding by using codes like [/] as described in the CHAT manual chapters 9 and 10.
6. Phonological fragments and fillers, as described in section 8.5 of the CHAT manual
7. If needed, you can add inserted or "back channel" communicators as described in section 9.10.2 of the CHAT manual. These may have been put onto their own line by the ASR and would be better represented through the inserted format.
8. Adding Gem markers as needed. For PsychosisBank, these could be simply @G: first, @G: second etc.
9. Add [+ exc] post-code for non-task utterances (only for picture description, Cinderella, PBJ; not for free speech tasks)

For a list of CHAT coding symbols, see the cheatsheet at
https://aphasia.talkbank.org/cheatsheet.doc

Useful editor commands:
F4 – plays current bulleted line, also command-click
F5 – redo bulleting
Esc-8 – continuous play
Esc-A – open/hide bullets