Batchalign can be downloaded from https://github.com/talkbank. It is a Python script, created by Houjun Liu, that uses either Whisper or Rev-AI automatic speech recognition (ASR) and the Unix version of CLAN to go from raw audio to a well-formatted CHAT transcription for analysis by CLAN and inclusion in TalkBank. Batchalign also has the capacity to use Universal Dependency taggers within the Stanza framework to add %mor and %gra morphological tiers to a CHAT transcript in dozens of languages.

Installation –

Batchalign is available through the Python Package Index. To get started, begin by installing Python. We support Python version 3.9, 3.10 but recommend version 3.11 strongly. Python 3.12 is *not supported*.

To set up Python 3.11, follow the instructions relevant to your platform:

MacOS

1. Install Homebrew via these instructions
2. Open a Terminal window and execute “brew install python@3.11” to install Python. You can find the macOS Terminal application in /Applications/Utilities. This terminal window is where you will enter Batchalign commands.
3. Open a new Terminal window, and execute “pip3 install -U batchalign”
4. If you don't already have pip3, the best way to get it is to install "brew" by going to brew.sh, copying the long install command, pasting that into your terminal window, and then waiting about 5 minutes as brew installs lots of useful things, including ffmpeg.

Windows

1. If you have a previous version (or Python 3.12) installed, its best to remove it. To do this, search for “Add or remove programs” using the Windows search bar, find other Python versions, tap the entries, and tap “Uninstall”.
2. Download and run the Windows Python 3.11 installation binary here.
3. Open a Command Line Window and execute “py -m pip3 install -U batchalign”. You can find the Command Line application by searching “Command Line” in the Windows search bar.

Linux

Follow your distribution’s instructions. Python 3.11 would be best. To install Batchalign:
pip3 install -U batchalign

Usage

Once installed, the usage of Batchalign follows these steps. These instructions supercede the earlier descriptions in our 2023 published article in JSLHR. Processing now runs all the way to the end, and you only do cleanup after it has completed.

1. At the level of ~/ or your home directory, you should create a folder using: mkdir ba_data. Then use cd ba_data to go inside that folder and create subfolders using: mkdir input and mkdir output.
2. When you run your first Batchalign command, you will be asked to run the setup. You will be asked if you want to insert a rev-ai key and it will give you instructions on how to get that. You will need to open a rev.ai account to transcribe speech. Rev-AI provides you with 6 free hours for your new account. After that, charges are $.02/minute of audio. Go to https://rev.ai , sign up, and on the left side of your dashboard, you will find a tab called Access Token. Click generate to generate a new token, copy and paste the key to somewhere you can find when using "batchalign setup". When signing up, you can configure your profile to have no data saved on their servers. Once you have provided your rev-ai key, it is saved in a file and you will not need to provide it again.
3. For **transcribe**, you just put media into input. For **morphotag**, you just put the transcript. For **align** you must put both media and transcript.
4. Batchalign works on .wav files. Other inputs are converted to .wav using ffmpeg. To install ffmpeg on Windows, follow these instructions.
5. If you record with iPhone, the format is m4a. Since batchalign and ffmpeg only accept mp3, mp4, and wav, you need to convert m4a to .wav. You can do this using an online converter site such as this one: https://cloudconvert.com/m4a-to-wav .
6. Batchalign supports three different processes with these three verbs:

- **transcribe** provides transcription directly from audio or video. This only requires raw media files (audio or video) in /input.
- **align** produces utterance- and word-level alignment of a bulleted text when you place both the media and transcript files into /input
- **morphotag** uses Stanza/UD to add %mor and %gra lines to a transcript. It does not require a media file.

7. You can put collections of files or even hierarchies of folders into the input folder and the output will preserve the herarchy. As it works, Batchalign will provide file-by-file feedback including either "Done" or "Fail" as well as warnings for segments that fail processing. However, progress counter doesn't increment while material is being processed by Rev-AI.

Cleanup

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

1. The outputs from **morphotag** and **align** should be checked using the CHECK and Chatter programs as described in the CLAN manual. CLAN can be downloaded from https://dali.talkbank.org
2. The output from **transcribe** will pass CHECK and Chatter, but it will need to be corrected in several ways.

3. To facilitate this process, it is best to first remove the %wor line using this CLAN command: trim -t%wor *.cha +1   You can add it back later in a second pass of **align**.
4. The next steps are done by opening each file in the CLAN Editor.
5. First, you will need to correct the 4-letter Speaker IDs created by Rev-AI from the form PAR0 or PAR1 to standard CHAT roles such as CHI for Target_Child, MOT for Mother, or INV for Investigator. Please consult the CHAT manual section 7.2 for the list of possible roles. You need to correct this both in the @Participants line and the @ID fields and later on throughout the transcript. You can get new correct @ID fields by deleting the old ones and then running escape-L to check the file which then automatically enters new ones based on the @Participants line.
6. Once you have the correct set of Participant IDs, you can choose "update" in the Tiers menu to get shortcuts for insertion of each ID as needed. Also, you can use CLAN's query-replace function (command-R) to change forms like *PAR0: to *CHI: throughout the file as needed.
7. A faster way of handling the two steps above is to use CLAN's ROLES program to automate the process. Following instructions in section 11.10 of the CLAN manual, you can create a roles.cut file and run this command: roles +croles.cut *.cha +1 to make all of these changes in one quick pass.
8. With the media file in the same folder as the transcript file, use CLAN's continuous playback command (Esc-8) to play back and listen through the whole transcript. This is the major job you need to do and it will take some time. As you go, you will want to fix three things.
    a. You can fix incorrect IDs using the shortcuts such as command-1 in the Tiers menu.
    b. You can correct incorrect words by retyping the correct forms.
    c. You can correct incorrect utterance segmentation by joining two utterances into one or breaking up an utterance into two pieces.
9. Once you are done with these fixes, you can run the file through the **align** version of Batchalign to restore the %wor lines.

Further Explanations:

1. **Examples**:

```
batchalign align ~/ba_data/input ~/ba_data/output
batchalign transcribe ~/ba_data/input ~/ba_data/output
batchalign transcribe ~/ba_data/input ~/ba_data/output --lang=spa
batchalign morphotag --lang=nld ~/ba_data/input ~/ba_data/output
```

2. **Languages**: To run the three basic commands (align, transcribe, morphotag) for languages other than English, you need to add the --lang switch, as in --lang=eng for working with English. The three-letter ISO-639 abbreviations can be found in the CHAT manual.

3. **Morphotag:** To tag transcripts for %mor and %gra using Universal Dependency models, you use commands in this form:

```
(batchalign) batchalign morphotag ~/ba_data/input ~/ba_data/output
```

The program reads the @Languages header in the CHAT files to determine which language model to use.