Nahuatl Corpus Overview and Workflow
(see CMU folder on Server)

Languages
1. Highland Puebla Nahuatl (high1278/azz)
2. Zacatlán-Ahuacatlán-Tepetzintla Nahuatl (zaca1241/nhi)

Metadata
See /cmu/Highland-Puebla-Metadata

Size
High1278: 954 audio files (plus one music file) > 200 hours
Zaca1241: 151 audio files (one channel field recordings during plant collection, average 2-3 minutes/file) < 6 hours

Location of audio
High1278 recordings delivered via Hard Disk to Shinji/Jiatong, some online in the CMU folder /cmu/Highland-Puebla-wav; with two speakers each is miked separately
Zaca1241 recordings on server at /cmu/Zacatlan-Nahuatl/Botanical-field-recordings; with two speakers one channel with speakers speaking in order, no overlap

Transcription status:
High1278:

- Almost all transcribed in Transcriber first (but very accurate draft)
- Approximately 300–350 of the 954 files have been proofed and translated to Spanish (in ELAN)
- Recordings delivered via Hard Disk to Shinji/Jiatong

Zaca1241

- No recordings have been transcribed.

Dictionary
High1278: Dictionary of approximately 8500 entries (uploaded to /cmu/Highland-Puebla-Nahuat/04_Dictionary This will be totally reviewed by 2022 but is in good shape now for meaning, grammatical categories, etc. A MyShoebox folder should be downloaded if Toolbox is used to view the dictionary.

Grammar
Rough grammatical sketch of 14 chapters at /cmu/Highland-Puebla-Nahuat/09_Grammar

Workflow and goals
Transcriptions: The Highland Puebla Nahuatl transcriptions were first done in Transcriber by native speakers, and then reviewed by them. Presently Amith is giving a final review to each and consulting with native speaker colleagues as necessary. The corrected Transcriber transcription is then imported into ELAN and given to one of two native speakers to translate into Spanish. Approximately 350 files have been so reviewed and translated. The goal is to finish the approximately 950 transciptions by the end of 2021. At that time all the material will be in ELAN with a transcription and translation. This corpus will be used for both ASR and NLP development.

Dictionary: Amith and a colleague, Ceferino Salgado, are presently reviewing the dictionary entries and structure. Approximately 100 entries are reviewed per week. Goal is to have 5,000 entries so proofed by December 2021 and finish the remainder by summer 2022. Nevertheless, the dictionary is presently quite useful.
Dictionary sound files: The goal is to have a sound file of each headword repeated 3 times. This should be done by mid 2022. The dictionary will also have 50,000 illustrative sentences (transcription, translation, and audio). About half of these sentences will be written by Amith and native speakers and then recorded by the same native speakers (reading from the text). The other half will be extracted from the corpus (i.e., natural conversation). Examples of the headword recordings, elaborated illustrated sentences, and corpus illustrated sentences are in
cmu/Highland-Puebla-Nahuat/04_Dictionary/Example-sound

Grammar:
In mid 2022 Amith will return to writing the corpus-based grammar of Highland Puebla Nahuatl. Approximately 50 chapters are planned.