About this resource:

The majority of the primary audio material of this deposit was gathered over a 10-month period from late May 2022 to early March 2023. It comprises 427 files totaling approximately 111 hours, 14 minutes, 27 seconds of recordings. Jonathan D. Amith (project director) recorded the 94 files from 25 to 30 May 2022. Amelia Domínguez Alcántara and Ceferino Salgado Castañeda recorded the remaining 333 files, from 1 June 2022 to 3 March 2023. In all cases Amith, Domínguez, and Salgado used a Sound Devices 722 digital recorder and Countryman e6 omnidirectional microphones. Most of these recordings are two-channel conversations, with each speaker on a separate channel, although a few (e.g., stories) are single-speaker and single-channel recordings. The remaining 151 recordings (totaling 3 hours, 1 minute, and 27 seconds) are all coded Tepet_BotFl indicating that they are field botany recordings from Tepetzintla. These were made by Osbel López-Francisco, a native Totonac speaker from Zongozotla, who was carrying out ethnobotanical fieldwork with Ceferino Salgado, a Nahuatl-speaker from Tacuapan, municipality of Cuetzalan del Progreso. He used a handheld Zoom H6n with a shotgun microphone for the plant collections 84091 to 84239, made between 21 June and 18 July 2019. Note that previously, in July 2016, Amith and Salgado collected 91 plants for which no field recordings were made. Note that one file recorded on 25 May 2022, the first, is not included as the speakers were Amelia Domínguez and Ceferino Salgado who were demonstrating for the native speakers of Omitlán the type of narrative that was desired. This file name is Xaltn_Narra_ADA300-CSC370_Historias-de-vida_2022-05-25-a.wav

The (ethno)botanical labels for 242 all plant collections in the municipality of Tepetzintla are included as reference (see pdf file named: Plant-Labels_Tepetzintla-Zacatlan-ethnobotanical-field-trips_2023-10-22.pdf). A reduced comma delimited csv file contains the metadata for collection number, family, scientific name, date collected, and name of person who identified the plant. As plants continue to be identified with their scientific names from the field photos taken, these files will be updated. Note again that only collections from 84091 to 84239 are accompanied by field recordings.

Please note that this initial OpenSLR deposit focuses on the audio corpus.

Please note that this initial OpenSLR deposit focuses on the audio corpus. Five future enhancements to this resource are envisioned at this present time: (1) Completed metadata, particularly a description of the content of each of the 578 recordings; (2) 10 hours of transcription by hand in ELAN, material that will provide the initial basis for transfer ASR; (3) A final deposit of the results of ASR transcriptions; (4) Corrections to the ASR transcriptions carried out by Amith and native speakers; (5) Reference to the ASR end2end recipe (GitHub) used to generate the ASR transcriptions.

The fieldwork for developing this corpus was supported by NSF Dynamic Language Infrastructure grant #2123578 entitled “Collaborative Research: Improving Techniques of Automatic Speech Recognition and Transfer Learning using Documentary Linguistic Corpora” (Jonathan D. Amith, PI). The speech processing facet of this research (Award #2123624) will be carried out by Shinji Watanabe (PI) and his team at Carnegie Mellon University.

Amith, Jonathan D., Amelia Domínguez Alcántara, Ceferino Salgado Castañeda, and Osbel López-Francisco, n.d., Audio corpus of Zacatlán-Ahuacatlán-Tepetzintla Nahuatl. Accessed [date] at https://www.openslr.org/.

Along with the audio recordings in .wav format (48KHz, 16-bit), at present this deposit includes the following files:

OpenSLR_Tepetzintla-Zacatlán-Nahuatl.pdf
(Document with information about this corpus)

Tepetzintla-Zacatlan-Nahuatl_Collaborators.txt
(List of all native speaker collaborators for this corpus)

Tepetzintla-Zacatlán-Nahuatl_File-list.txt
Tepetzintla-Zacatlán-Nahuatl_File-list.pdf
(list of all filenames with duration)

Plant-collections_Tepetzintla.csv
(list of all plant collections with collection number, family, scientific name, date collected, name of person who identified the plant)

Plant-Labels_Tepetzintla-Zacatlan-ethnobotanical-field-trips_2023-10-22.pdf
(labels for the 242 plant collections in the municipality of Tepetzintla, the audio for field recordings of the final 151 collections is included in this corpus)