About this resource:

The substantive material of this deposit was gathered over a 13-month period from February 2022 to March 2023. It comprised 657 files totaling approximately 119 hours, 26 minutes, 59 seconds of material. All but 81 files (i.e., 576 files) were recorded by Jonathan D. Amith (Project director) or Amelia Domínguez Alcántara and Ceferino Salgado Castañeda using 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 81 recordings, totaling 14 hours, 31 minutes, 59 seconds, are all coded Teotz_BotFl or Ixpal_BotFl. These were made by Mariano Gorostiza Salazar and Miriam Jiménez Chimil during a short trip (7 March to 16 March 2023) to photograph plants that have names in the Nahuatl spoken in the region of Tequila and Orizaba, Veracruz. The (ethno)botanical labels for these 81 plant observations are included as reference in this OpenSRL resource (see pdf file named: Plant-Labels_Tequila-Orizaba-ethnobotanical-field-trip_2023-10-22.pdf). As plants continue to be identified with their scientific names from the field photos taken, this file will be updated.

Fieldwork was coordinated locally by Gabriela Citlahua Zepahua, who also participated as a speaker in some of the recordings. Citlahua Zepahua was responsible for contacting the native speakers who generously participated in this research.

Please note that this initial OpenSLR deposit focuses on the audio corpus. Five future enhancements to the metadata for this corpus are envisioned at this present time: (1) Completed metadata, particularly a description of the content of each recording; (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 by Amith and native speakers of Orizaba Nahuatl; (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.

Please cite or use any material as follows (Corresponding author is Jonathan D. Amith jonamith@gmail.com).

Amith, Jonathan D., Amelia Domínguez Alcántara, Ceferino Salgado Castañeda, Gabriela Citlahua Zepahua, Mariano Gorostiza Salazar, and Miriam Jiménez Chimil, n.d., Audio corpus of Orizaba Nahuatl. Accessed [date] at https://www.openslr.org/.

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

OpenSLR_Veracruz-Orizaba-Nahuatl.pdf
(Document with information about this corpus)

Veracruz-Orizaba-Nahuat_Collaborators.txt

(List of all native speaker collaborators for this corpus)

Veracruz-Orizaba-Nahuatl_File-list.txt
(list of all filenames with duration)

Plant-observations_Veracruz.csv
(list of all plant observations with observation number, family, scientific name, date collected, name of person who identified the plant)

Plant-Labels_Tequila-Orizaba-ethnobotanical-field-trip_2023-10-22.pdf
(labels for the 81 plant observations the audio of which is included in this corpus)