Resources for end-to-end automatic speech recognition of Sierra Nororiental and Norte de Puebla Nahuat(l)

Highland Puebla Nahuat	954 files	Glottocode high1278	ISO 639-8 azz
Zacatlán-Ahuacatlán-Tepetzintla Nahuatl:	151 files	Glottocode zaca1241;	ISO 639-3 nhi

Corresponding authors: Jonathan D. Amith (jonamith@gmail.com)
Jiatong Shi (jshi34@jhu.edu)

Background, support, and licensing

Background

The substantive material of this deposit was gathered over ten years by Jonathan D. Amith (PI) and a team of native speaker colleagues who have participated in the project for many years, one from its inception in 2009. The speakers are:

Amelia Domínguez Alcántara: From Xaltipan, municipality of Cuetzalan del Progreso; born 1976
Ceferino Salgado Castañeda: From Tacuapan, municipality of Cuetzalan del Progreso; born 1984
Hermelindo Salazar Osollo: From San Miguel Tzinacapan, municipality of Cuetzalan del Progreso; born 1954

Amelia Domínguez has been with the project since its inception. Eleuterio Gorostiza Salazar was also an initial member of the research team but left to pursue a master's degree in linguistics

Eleuterio Gorostiza Salazar: From San Miguel Tzinacapan, municipality of Cuetzalan del Progreso; born 1978

Grant support

The following grants supported research that produced the primary material deposited here

NSF, Documenting Endangered Languages (Award #BCS-1401178), A Biological Approach to Documenting Traditional Ecological Knowledge in Synchronic and Diachronic Perspectives
NEH, Preservation and Access (Award #PD-50031-14), A Biological Approach to Documenting Traditional Ecological Knowledge in Synchronic and Diachronic Perspectives
Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (Award ME010), Floristics, Biodiversity, and Traditional Ecological Knowledge in the Sierra Nororiental of Puebla, Mexico
Endangered Language Documentation Programme, School of Oriental and African Studies (Award MDP0272), Documentation of Nahuat Knowledge of Natural History, Material Culture, and Ecology in the Municipality of Cuetzalan, Puebla.
NSF, Documenting Endangered Languages (Award #0756536), Nahuatl Language Documentation Project: Sierra Norte de Puebla. National Science Foundation, Documenting Endangered Languages ($291,798, Award #0756536)

Licensing

All material is made available under the Creative Common license CC BY-NC-SA (Attribution-NonCommercial-ShareAlike). Please cite or use any material as follows.

Citation

If only the recordings and transcriptions:

Amith, Jonathan D., Amelia Domínguez Alcántara, Hermelindo Salazar Osollo, Ceferino Salgado Castañeda, and Eleuterio Gorostiza Salazar, n.d., Audio corpus of Sierra Nororiental and Sierra Norte de Puebla Nahuat(l) with accompanying time-code transcriptions in ELAN.

If the ASR programming is cited:

Shi, Jiatong and Jonathan D. Amith. Automatic speech recognition of Sierra Nororiental and Sierra Norte de Puebla Nahuat(l).

Primary data from the municipality of Cuetzalan del Progreso (recordings and transcriptions)
Together the four speakers transcribed the audio of 954 recordings (see metadata) in Transcriber. All 954 Transcriber files are included with this deposit. Over the years Amith has repeated edited the original orthography as necessary, often using regular expression searches throughout the entire corpus to ensure consistency.

Of these 954 Transcriber transcriptions (see folder Transcriber-files-all-954) a total of 738 were imported into ELAN for the review (proofing by Amith) and translation into Spanish, by A. Domínguez, C. Salgado, and H. Salazar). At present these ELAN files are divided into two sets

- Set 1: These are 299 ELAN files that Amith has carefully reviewed, listening to the recordings while editing the transcriptions as necessary. The vast majority have been freely translated into Spanish.
- Set 2: These are 439 ELAN files that Amith will review over the coming year, after which they will be translated by the three native speaker colleagues. This task should be finished by late Spring or early Summer 2021. Note that once this is done the remaining 216 files (of the total 954) will be

All recordings for more with two speakers used separate microphones for each speaker. See Metadata for further information.

The 954 recordings can be loosely grouped under the following topics. See the Metadata file for a detailed description of the content of each recordings. Note that the botany recordings are mostly about specific species. The scientific (and Indigenous) names for these species is contained in the metadata for these recordings.

Genre of recording	Number of recordings
Agriculture	10
Botany	579
Hunting and fishing	18
Food	20
Beliefs	4
Stories	22
Material culture	44
Medicine	79
Narrations and life histories	62
Ritual	3
Traditions	12
Zoology	101
TOTAL	954

Primary data from the municpality of Tepetzintla (recordings)
As part of a research project on comparative ethnobotany in Nahuatl-, Totonac-, and Mixtec-speaking communities Amith, Salgado Castañeda and Osbel López Francisco (a botanist and Totonac speaker from Zongozotla, Puebla) have begun research in the municipality of Tepetzintla (19.96701, -97.84082) in which there is one Totonac-speaking community (Tonalixco) whereas the rest speak Nahuatl. The only recordings to date from this municipality are recordings made with a handheld (internal microphone) Zoom H4n recorder. All 151 recordings were made at the time plants were collected with two Nahuatl speakers (a woman, Josefa Fernández, and her daughter, María Concepción Robles Fernández (see metadata). The recordings are often of poor quality given the inexperience of the individuals who made the recordings and manifest some clipping and, at other times, low signal-to-noise ratios. The recordings have been normalized which explains why in some cases there is clipping even though the dynamic range does not reach the maximum.
All recordings were made at the time of collection of a particular plant. The filename of the recording references the collection number and the best opinion of the family and genus of the plant at the time it was collected. Again,

the metadata gives more precise information, including the plant name and a description of the contents of the recording.

Note that the Nahuatl from the municipality of Tepetzintla is particularly interesting as a challenge to extending coverage of the ASR tools developed for Nahuat from the municipality of Cuetzalan, particularly because of its different phololology.

- Tepetzintla has the [λ] (<tl>) common to most Nahuatl languages, a sound that is reflexed as [t] in Cuetzalan (cf. titla:katl 'you are a man' vs. tita:kat in Cuetzalan.
- Tepetzintla has an implosive voiced bilabial stop [ɓ] where other Nahuat(l) languages have [k^w] (or at times [k]) (cf. ɓowitl vs. k^wawit o kowit. This implosive is virtually undocumented in Nahuatl languages
- Tepetzintla has word final consonant clusters, perhaps an influence from neighboring Totonac phonotactics: cf. Tepetzintla witstl [] vs. Cuetzalan witsti, 'thorn'.

Additional data files uploaded to OpenSLR

In addition to the 954 audio files, corresponding 954 Transcriber files, and the 738 ELAN files (299 finalized and mostly translated), pertinent to Sierra Nororiental de Puebla Nahuat (high1278) and the 151 audio files (no transcriptions) pertinent to Zacatlán-Ahuactlán-Tepetzintla Nahuatl (zaca1241) the following files have been uploaded to Openslr

Summary file lists (tab delimited list of all filenames, unique identifiers, and duration)

Summary-data_Cuetzalan-954-recordings_Tab-delimited.txt
Summary-data_Tepetzintla-151-Field-recordings_Tab-delimited.txt

Metadata: Complete

Metadata_ Cuetzalan-954-recordings.xml
Metadata_Tepetzintla-151-Field-recordings.xml
Metadata_Persons-who-participated-by-recording.xml