

## Resources for end-to-end automatic speech recognition of Yoloxóchitl Mixtec

The substantive material of Totonac from the northern sierras of Puebla and adjacent areas of Veracruz were compiled starting in 2016 by Jonathan D. Amith and continue to the present as part of a joint effort by Amith and Osbel López Francisco, a native speaker biologist from Zongozotla. The corpus of materials comprises the following sets of recordings.

### Narratives and conversations: 105 total files

Community	Glottocode/ ISO 639-3	Municipality	Decimal coordinates	Number of files	Total duration hours:minutes:seconds
Ecatlán	high1243/tos	Jonotla	20.05224, -97.55677	5 files	00:56:41
Nanacatlán	high1243/tos	Zapotitlán de Méndez	20.00354, -97.67378	12 files	00:57:49
Tonalixco	high1243/tos	Tepetzintla	19.98176, -97.84053	1 file	00:06:49
Zongozotla	high1243/tos	Zongozotla	19.97857, -97.72693	87 files	07:03:57
<b>Total</b>				105 files	09:05:16

#### File structure (Deposit: March 2021)

- *Audio*
  - *Ecatlán* (5)
  - *Nanacatlán* (12)
  - *Tonalixco* (1)
  - *Zongozotla* (87)
- *Transcription*
  - *Ecatlán* (0)
  - *Nanacatlán* (0)
  - *Tonalixco* (0)
  - *Zongozotla* (87)

#### Recordings and transcriptions

Recordings from Ecatlán, Tonalixco, and Zongozotla were made by J. Amith with a Marantz PMD671 digital field recorder and either a pair of Shure SM-10a or Countryman e6 earworn mics. Nanacatlán recordings were made by Osbel López with a handheld Zoom H4n. All metadata is in a accompany XML file.

All transcriptions were made by Osbel López. To date (March 2021) only the 87 Zongozotla files have been transcribed. By May 2021 the remaining 18 files (02:01:17) will be transcribed. Transcriptions were initially done in Transcriber and imported into ELAN with separate tiers for each speaker. Translations were then made in ELAN in separate tiers, one for each speaker. Most recordings have two speakers in conversation though with little overlap in speech.

**Notes on transcription:** In the transcriptions distinctive Totonac vowel length is marked by a colon. Voiceless velar and uvular stops are represented by <k> and <q>, respectively. The voiceless lateral fricative <lh> contrasts with the voiced lateral approximant <l>. Vowel laryngealization (creaky voice) is often realized as a glottal stop (particularly in word final position). The accurate representation of laryngealization is still problematic and as greater skill is developed in distinguishing this sound the transcriptions will be updated. Note that transcription filenames end with the date of the most recent edition.

**Filenames:** These follow a standard format

Zongo\_Botan\_SLC388-IPN389\_Acanthaceae-Justicia-aurea\_2018-07-26-j\_ed-2019-08-19.eaf

Village-Genre-Left.channel.speaker.code-right.channel.spk.code\_Theme\_UID/often=date.recorded\_edition.date  
For speaker codes and demographics, please see Appendix at end of this document.

### Ethnobotanical field recordings: 541 total files (deposit to openSLR planned for summer 2021)

Community		Municipality	Decimal coordinates	Number of files	Total duration hours:minutes:seconds
Coahuatlán	filo1235/ tjp (provisional)	Coahuatlán, Veracruz	20.25681, -97.73054	53 files	02:35:21
Ecatlán	high1243/tos	Jonotla	20.05224, -97.55677	34 files	01:42:05
Mecatlán	filo1235/ tjp (provisional)	Mecatlán, Veracruz	20.19536, -97.66564	187 files	07:25:01
Nanacatlán	high1243/tos	Zapotitlán de Méndez	20.00354, -97.67378	116 files	02:00:30
Tonalixco	high1243/tos	Tepetzintla	19.98176, -97.84053	151 files	03:26:51
<b>Total</b>				541 files	17:09:48

#### Recordings and transcriptions

Ethnobotanical field recordings were made with a handheld Zoom H4n field recorder (set at 48KHz, 16-bit) with recordings made at the time a plant was collected by interviewing local native speaker consultants one after the other (no overlapping speech) in a one-channel recording. The duration of most was short, from 2 to 4 minutes (see metadata for precise information).

To date (March 2021) none of these 541 files have been transcribed although all have been described in detail in regards to their content. The recordings and metadata will be uploaded to OpenSLR during the summer of 2021. Note that because of lack of experience in some recordings occasionally significant clipping did occur.

#### Plant name elicitation: > 2000 short files (deposit to openSLR planned for summer 2021)

Plant names have been elicited for Zongozotla, Atlequizayán, Ecatlán, and Tonalixco. Recordings are in general three repetitions of the plant name in isolation. In some cases a fourth repetition was elicited embedded in words preceding and following the plant name. This was particularly important as in many Totonac languages phrase final open syllables lose the final vowel.

As single word utterances, even a repeated series of tokens, is not particularly valuable for ASR, these files have not been uploaded to OpenSLR though they will in the near future.

#### Grant support

Production of the corpus was generously supported by the National Science Foundation, Documentation Endangered Languages program, the Endangered Language Documentation Programme (ELDP) at the School of Oriental and African Studies, and the Jacobs Research fund:

- 2020 Comparative Totonacan Ethnobotany: Documentation of the Nomenclature, Classification, and Use in Three Communities (Tonalixco, Ecatlán, Pisaflores); Jacobs Research Fund
- 2019 Community-based Ethnobotany in the Sierra Nororiental de Puebla: Zongozotla and Tonalixco Totonac; Jacobs Research Fund
- 2016–17 Totonac ethnobotanical knowledge: Documenting traditional ecological knowledge across communities. Endangered Language Documentation Programme, School of Oriental and African Studies, University of London (MDP0352: David Beck, PI; Jonathan D. Amith, co-PI)
- 2014–21 A Biological Approach to Documenting Traditional Ecological Knowledge in Synchronic and Diachronic Perspectives, National Science Foundation, Documenting Endangered Languages and Anthropology (Award #BCS-1401178), including two supplement : Award#1646724, Award#2039336
- 2014–18 A Biological Approach to Documenting Traditional Ecological Knowledge in Synchronic and Diachronic Perspectives, National Endowment for the Humanities (Award #PD-50031-14)

#### License

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 the recordings and transcriptions are cited in general, please use:

Amith, Jonathan D., and Osbel López Francisco. n.d. Audio corpus of Totonac recordings from northern Puebla and adjacent areas of Veracruz. OpenSLR##.

If a particular recording is used please cite the speakers and filename followed by In Jonathan D. Amith and Osbel López Francisco. Audio corpus of Totonac recordings from northern Puebla and adjacent areas of Veracruz. OpenSLR## accessed yyyy-mm-dd.

Corresponding author is [jonamith@gmail.com](mailto:jonamith@gmail.com)

## Appendix: Speaker codes and demographics

\last\_name López Francisco  
\first\_name Osbel  
\uid 385  
\code OLF385  
\sex Male  
\village Zongozotla  
\mpio Zongozotla  
\estado Puebla  
\age  
\birthapprox  
\birthyear 1991  
\notes Began work with Canek Ledesma for botanical field collection in Totonac village of Zongozotla. Later hired to document Totonac plant nomenclature.  
\type Totonac native speaker consultant

\last\_name López Cano  
\first\_name Sebastián  
\uid 388  
\code SLC388  
\sex Male  
\village Zongozotla  
\mpio Zongozotla  
\estado Puebla  
\age  
\birthapprox  
\birthyear 1968  
\notes Father of Osbel López Francisco  
\type Totonac consultant

\last\_name Pérez Núñez  
\first\_name Isac  
\uid 389  
\code IPN389  
\sex Male  
\village Zongozotla  
\mpio Zongozotla  
\estado Puebla  
\age  
\birthapprox  
\birthyear 1971  
\notes Father-in-law of Osbel López Francisco  
\type Totonac consultant

\last\_name Andrés Hernández  
\first\_name Crisóforo  
\uid 390  
\code CAH390  
\sex Male  
\village Mecatlán  
\mpio Mecatlán  
\estado Veracruz

\age  
\birthapprox  
\birthyear 1982  
\notes  
\type Totonac consultant

\last\_name Andrés Cruz  
\first\_name Eleuterio  
\uid 391  
\code EAC391  
\sex Male  
\village Mecatlán  
\mpio Mecatlán  
\estado Veracruz  
\age  
\birthapprox  
\birthyear 1942  
\notes  
\type Totonac consultant

\last\_name Luciano Ortega  
\first\_name Marcelino  
\uid 392  
\code MLO392  
\sex Male  
\village Mecatlán  
\mpio Mecatlán  
\estado Veracruz  
\age  
\birthapprox 1955  
\birthyear  
\notes  
\type Consultant

\last\_name López Bonilla  
\first\_name Miguel Juan  
\uid 400  
\code MJL400  
\sex Male  
\village Tonalixco  
\mpio Tepetzintla  
\estado Puebla  
\age  
\birthapprox  
\birthyear 1941  
\notes  
\type

\last\_name López Alemán  
\first\_name Andrés  
\uid 403  
\code ALA403  
\sex Male

\village Tonalixco  
\mpio Tepetzintla  
\estado Puebla  
\age  
\birthapprox  
\birthyear 1976  
\notes Hijo de Miguel Juan López Bonilla.  
\type

\last\_name Orozco Bautista  
\first\_name Carlos  
\uid 409  
\code COB409  
\sex Male  
\village Ecatlán  
\mpio Jonotla  
\estado Puebla  
\age  
\birthapprox  
\birthyear 1963  
\notes  
\type Totonac botanical consultant

\last\_name Orozco Desión  
\first\_name Evencio  
\uid 410  
\code EOD410  
\sex Male  
\village Ecatlán  
\mpio Jonotla  
\estado Puebla  
\age  
\birthapprox  
\birthyear 1941  
\notes  
\type Totonac botanical consultant

\last\_name Pacheco Cruz  
\first\_name Agustina María  
\uid 415  
\code APC415  
\sex Female  
\village Tonalixco  
\mpio Tepetzintla  
\estado Puebla  
\age  
\birthapprox  
\birthyear 1986  
\notes  
\type

\last\_name Gómez Hernández  
\first\_name Nicolasa

\uid 416  
\code NGH416  
\sex Female  
\village Tonalixco  
\mpio Tepetzintla  
\estado Puebla  
\age  
\birthapprox  
\birthyear 1981  
\notes  
\type

\last\_name Valle Aragón  
\first\_name María Dolores  
\uid 417  
\code MDV417  
\sex Female  
\village Tonalixco  
\mpio Tepetzintla  
\estado Puebla  
\age  
\birthapprox  
\birthyear 1994  
\notes  
\type

\last\_name Galindo Bautista  
\first\_name Benjamín  
\uid 419  
\code BGB419  
\sex Male  
\village Ecatlán  
\mpio Jonotla  
\estado Puebla  
\age  
\birthapprox 1945  
\birthyear  
\notes No es seguro que su segundo apellido es Bautista.  
\type

\last\_name Galindo Bautista  
\first\_name Aristeo  
\uid 420  
\code AGB420  
\sex Male  
\village Ecatlán  
\mpio Jonotla  
\estado Puebla  
\age  
\birthapprox  
\birthyear 1940  
\notes  
\type Totonac consultant for botanical collections.

\last\_name Bautisa Nicolás  
\first\_name Angelina  
\uid 421  
\code ABN421  
\sex Female  
\village Ecatlán  
\mpio Jonotla  
\estado Puebla  
\age  
\birthapprox  
\birthyear 1936  
\notes  
\type Totonac consultant for botanical collections.

\last\_name Desión Rodríguez  
\first\_name Adalberto  
\uid 422  
\code ADR422  
\sex Male  
\village Ecatlán  
\mpio Jonotla  
\estado Puebla  
\age  
\birthapprox  
\birthyear  
\notes  
\type Totonac consultant for botanical collections.

\last\_name Dionicio Orozco  
\first\_name Valencia  
\uid 423  
\code VDO423  
\sex Female  
\village Ecatlán  
\mpio Jonotla  
\estado Puebla  
\age  
\birthapprox  
\birthyear  
\notes  
\type Totonac consultant for botanical collections.

\last\_name García Ramírez  
\first\_name Gustavo  
\uid 424  
\code GGR424  
\sex Male  
\village Ecatlán  
\mpio Jonotla  
\estado Puebla  
\age  
\birthapprox



\birthyear  
\notes  
\type Totonac consultant for botanical collections.

\last\_name Orozco Valencia  
\first\_name Dionisio  
\uid 425  
\code DOV425  
\sex Male  
\village Ecatlán  
\mpio Jonotla  
\estado Puebla  
\age  
\birthapprox  
\birthyear  
\notes  
\type Totonac consultant for botanical collections.

\last\_name Ponce García  
\first\_name Areli  
\uid 426  
\code APG426  
\sex Female  
\village Zongozotla  
\mpio Zongozotla  
\estado Puebla  
\age  
\birthapprox  
\birthyear  
\notes  
\type

\last\_name Sahinos Pérez  
\first\_name Epifania  
\uid 427  
\code ESP427  
\sex Female  
\village Zongozotla  
\mpio Zongozotla  
\estado Puebla  
\age  
\birthapprox  
\birthyear  
\notes  
\type

\last\_name Luna Ramos  
\first\_name Amador  
\uid 428  
\code ALR428  
\sex Male  
\village Nanacatlán  
\mpio Zapotitlán de Méndez

\estado Puebla  
\age  
\birthapprox  
\birthyear 1947  
\notes  
\type Asesor en Totonaco, particularmente acerca de plantas.

\last\_name López Toribio  
\first\_name Natividad  
\uid 429  
\code NLT429  
\sex Female  
\village Nanacatlán  
\mpio Zapotitlán de Méndez  
\estado Puebla  
\age  
\birthapprox  
\birthyear 1952  
\notes  
\type Asesora en Totonaco, particularmente acerca de plantas.

\last\_name Valencia Zenaida  
\first\_name Valencia  
\uid 430  
\code VVZ430  
\sex Female  
\village Ecatlán  
\mpio Jonotla  
\estado Puebla  
\age  
\birthapprox  
\birthyear  
\notes Confirmar nombre y consigue año de nacimiento  
\type

\last\_name Bautista Nicolás  
\first\_name María Inés  
\uid 431  
\code IBN431  
\sex Female  
\village Ecatlán  
\mpio Jonotla  
\estado Puebla  
\age  
\birthapprox  
\birthyear 1941  
\notes  
\type

\last\_name Pérez  
\first\_name Luis  
\uid 432  
\code LPx432

\sex Male  
\village Ecatlán  
\mpio Jonotla  
\estado Puebla  
\age  
\birthapprox  
\birthyear  
\notes Conseguir datos de nacimiento y segundo apellido.  
\type

\last\_name Hernández Pérez  
\first\_name Manuel  
\uid 441  
\code MHP441  
\sex Male  
\village Zongozotla  
\mpio Zongozotla  
\estado Puebla  
\age  
\birthapprox 1950  
\birthyear  
\notes Asesor de colectas de plantas.  
\type

\last\_name García Ponce  
\first\_name Miguel  
\uid 442  
\code MGP442  
\sex Male  
\village Zongozotla  
\mpio Zongozotla  
\estado Puebla  
\age  
\birthapprox  
\birthyear 1987  
\notes Asesor de colectas de plantas.  
\type

\last\_name Francisco Gaona  
\first\_name Avelina  
\uid 443  
\code AFG443  
\sex Female  
\village Zongozotla  
\mpio Zongozotla  
\estado Puebla  
\age  
\birthapprox  
\birthyear 1970  
\notes Gave a small amount of informatino (one plant?) for collections.  
\type

\last\_name Orozco Desión

\first\_name Evencio  
\uid 444  
\code AOD444  
\sex Male  
\village Ecatlán  
\mpio Jonotla  
\estado Puebla  
\age  
\birthapprox  
\birthyear 1941  
\notes  
\type Consultant on Ecatlán ethnobotany.

\last\_name Bautista Castillo  
\first\_name Raúl  
\uid 451  
\code RBC451  
\sex Male  
\village Ecatlán  
\mpio Jonotla  
\estado Puebla  
\age  
\birthapprox  
\birthyear 1957  
\notes Fue asesor en grabar nombres de plantas con David Beck y Osbel López Francisco. También fue entrevistado y grabado acerca de varios puntos en cuanto al léxico, morfología y gramática del totonaco de Ecatlán.  
\type