Ethnobotanical research and language documentation of Nahuatl

By Jonathan D. Amith, Amelia Domínguez Alcántara, Hermelindo Salazar Osollo, Ceferino Salgado Castañeda, and Eleuterio Gorostiza Salgado (2021)

Contents

1. Project history

The present corpus represents the results of Amith's interest in the Nahuat spoken in the municipality of Cuetzalan del Progreso, located in the Sierra Nororiental de Puebla (Glottocode = high1278; ISO 639-3 code = azz) dates to 2006, when he first visited the region and then returned a year later, late summer 2007, to give a two-week seminar on the Náhuat(l) language in the installations of the Indigenous collective Tosepan Titataniske. He noticed the unusual interest and aptitude of many of the attendees and towards the end of the workshop proposed working together on language documentation and asked the collective to collaborate. That fall Amith successfully applied for a grant from the National Science Foundation, Documenting Endangered Languages. The first award he received was from NSF-DEL (Award #0756536); he has continued to work in the region, combining a focus on Nahuat(l) with research on Yoloxóchitl Mixtec (Glottocode = yolo1241; ISO 639-3 = xty) and, more recently several Totonac languages from the Sierra Nororiental and Sierra Norte de Puebla, particularly the Totonac of Zongozotla (Glottocode = high1243; ISO 639-3 = tos).

After receiving the initial financing Amith worked with an original team of three native speakers all of whom he had met in the workshop the previous year: Amelia Domínguez Alcántara from Xaltipan; Eleuterio Gorostiza Salazar from San Miguel Tizacapan; and Aldegundo González Álvarez who worked half-time on the project and as a member of the Tosepan Titataniske council helped articulate collaboration between the language documentation effort and the collective. This team stayed together for several years before González left to attend to his obligations with Tosepan Titataniske full-time. He was replaced by Hermelindo Salazar Osollo (from San Miguel Tzinacapan), a colleague of Domínguez Alcántara from when they both worked in adult education, and the uncle of Eleuterio Gorostiza Salazar. A few years later Gorostiza Salazar left the project to pursue a master's degree in linguistics in Lingüística Indoamericana at CIESAS (Centro de Investigación y Estudios Superiores en Antropología Social) in Mexico City. Soon thereafter Ceferino Salgado Castañeda joined the team; he had previously worked as a consultant with the linguist Una Canger on Sierra Nororiental de Puebla Nahuat.

2. Project materials and data

a. Dictionary: The early years of research were dedicated to developing an orthography and familiarizing all team members with the writing system and, particularly, in recognizing phonemic vowel length. To ensure consistency in transcriptions the team worked for years on a dictionary which now (part of this deposit) comprises 8,704 entries and some 4,400 manuscript pages. Most of the early decisions on accurate phonemic representations of Nahuat words have endured.

b. Grammar: Alongside the dictionary development Amith and the research team explored the basics of Nahuat morphology and syntax. The thirteen chapters included in this deposit represent the results of this effort.

c. Corpus: A third basic and early concern of language documentation was the development of a digitally recorded, transcribed and translated corpus. At present the corpus comprises 954 audio recordings of speech (185 hours 20 minutes) and 1 short recording of music. Recently 15 mp4 files were added totalling 4 hours 25 minutes. All 200 hours of material has been carefully transcribed. Approximately one-third of the corpus has been translated into Spanish in ELAN. The audio files are found in \Mpio-Cuetzalan-del-Progreso\02_Multimedia\Audio whereas the video files are in \Mpio-Cuetzalan-del-Progreso\02_Multimedia\Video. The Transcriber and ELAN transcriptions of the video are located in the same folder as the .mp4 recordings. The Transcriber and ELAN transcriptions of the audio are in a separate folder: \Mpio-Cuetzalan-del-Progreso\03_Transcriptions. The ELAN_Finalized subfolder includes all 299 translated transcrptions. The ELAN_Not-Finalized subfolder includes two sets of ELAN files: Botany (279 files) and Not-Botany (160) files. These 439 ELAN files will be translated over the following year and be available in the subsequent edition. Together the Not-Finalized ELAN files (439 total) and the Finalized files (299, translated) total 738 files. The remaining (954-738) 216 transcribed files are presently only in Transcriber.

As part of the recent extension of research into comparative Nahuat(l) and Totonac ethnobiology, an additional corpus of 151 field recordings were in the municipality of Tepetzitlan made with a handheld Zoom H4n recorder.

Summary of multimedia files

Highland Puebla Nahuat *Glottocode high1278* *ISO 639-8 azz*	955 audio files 15 video files	185 hours 25 mins. 4 hours 25 mins.	
Zacatlán-Ahuacatlán-Tepetzintla Nahuatl: *Glottocode zaca1241* *ISO 639-3 nhi*	151 audio files	3 hours 1 min.	

Metadata for the corpus recordings in the municipalities of Cuetzalan and Tepetzintla is provided in the files found in the \01_Metadata_and_Contents for each municipalitiy. There is no corresponding metadata for Hueyapan and Huitzilan de Serdán because as of now there are still no multimedia recordings from these latter two municipalities.

But in addition to the csv and tab-delimited metadata files, Amith has also generated two summary catalogues for both Cuetzalan (see \docs\08_Catalogue-of-multimedia-Sierra-Nororiental-de-Puebla.pdf) and Tepetzintla (see \docs\ 09_Catalogue-of-multimedia-Nahuatl-in-Municipality-of-Tepetzintla.pdf). These catalogues are provided to facilitate discovery of material by potential users and stakeholders.

d. 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 not only has a distinct phonology, morphology and syntax from that spoken in the municipality of Cuetzalan but it has a phonology that is quite distinct from any other documented Nahuatl language

- 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'.

3. Ethnobotanical and botanical focus

a. Ethnobotany: In December 2007 Amith collected his first plant in the municipality of Cuetzalan, a plant called a:kwitaxo:chitl (Justicia aurea Schltdl.). Since then he has collected and photographed a total of 6,616 plants as part of an extensive study of comparative plant nomenclature, classification, and economic and symbolic use in Nahuat(l)-speaking villages of the northern sierra of Puebla. A total of 5755 plants were collected in Nahuat(l)-speaking communities. The remaining 861 were collected in Totonac-speaking communities (Zongozotla [574], Atlequizayán [283], and Ecatlán [4]. These 861 collections are only part of the total number of plants (2,920) collected in eleven Totonac communites; these Totonac collections along with documentation (textual and audio) of plant names is included in another deposit.

It is clear that a major focus of the recordings is plant nomenclature, classification, and use. For example, the following table lists the "genre" of the 955 audio recordings from the municipality of Cuetzalan. The 955 recordings can be loosely grouped under the following topics. Specific discussionis about plants comprise 579 recordings (just over 60 % of the total). The botany recordings are mostly about specific species. The scientific (and Indigenous) names for these species is contained in the metadata for these recordings. Many of the other genres are also pertinent to plant knowledge. For example Material Culture recordings discuss the use of plants in making objects of daily use.

Genre of Nahuat language recordings from the municpality of Cuetzalan

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
Music	1
Narrations and life histories	62
Ritual	3
Traditions	12
Zoology	101
TOTAL	955

b. Botany

Although endangered language documentation is primarily linguistic, this project has enjoyed the collaboration and support of many botanists and herbaria. The documentation of plant nomenclature, classification, and use in Indigenous communities was accompanied by extensive floristic studies in these communities. The result was the 6,616 plants collected in this "Nahuat(l)" study, including several species new to science and over 200 new state registers. The botanical data is all included in this deposit both metadata and pdf files of herbarium labels. Morever, this deposit includes 7,543 high quality photographs of plants in situ, all of which are linked to collection data and identifications, most to species. The photographs are offered to help other researchers who might want to use them as elicitation tools in their own ethnobotanical research, particularly if it is carried out in the same or similar regions.

4. 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)

5. Future plans and editions

All of the material deposited is in constant review and edition. The grammar, in fact, is in a very preliminary form, and the dictionary will need several years of work (including, besides careful editing of senses and subentries, the linking if sound files of headwords and illustrative sentences) before it is close to publishable. But these materials are all included both for "security" (to ensure that they are available to future generations) and to enable their use while the project continues. The following are some year-by-year plans for future editions of this deposit.

+ 1 Year

- An additional 400–450 audio recordings from Cuetzalan will be carefullly proofed and translated.
- The dictionary will be carefully edited and approximately one-third (3,000) of the entries will be completed.
- Dictionary headword recordings will be made for the aforementioned 3,000 entries
- The illustrated botanical field guide for Cuetzalan will be completed and pending plant identifications made. The field guides for Hueyapan, Huitzilan, and Tepetzitla will also be developed.

+ 2 Years

- The audio recordings from Cuetzalan will be finished: proofed and translated
- Amith will try to work with computational linguists to fully parse and gloss the Nahuat transcriptions
- The second third of the dictionary will be edited (up to entry 6,000)
- Dictionary headword recordings will be made for the aforementioned 3,000 entries
- A comparative encyclopedia of Nahuat(l) and Totonac botanical knowledge will be drafted

+ 3 Years

- The final third of the dictionary will be edited (up to entry 9,000)
- Dictionary headword recordings will be made for the final one-third of the dictionary

- The audio recordings will be reviewed for illustrative sentences of word use in context. The goal will be to develop 4,000 such illustrative sentences and link them to the dictionary
- The research team will elaborate and record illustrative sentences that will be linked to dictionary entries and senses
- Work on revising the reference grammar will begin.

6. Directory structure

The file structure follows the basic concept of LDC guidelines as there is a \data \docs folder below the root directory. In regard to dtd see the discussion in section 8 below (Data formats)

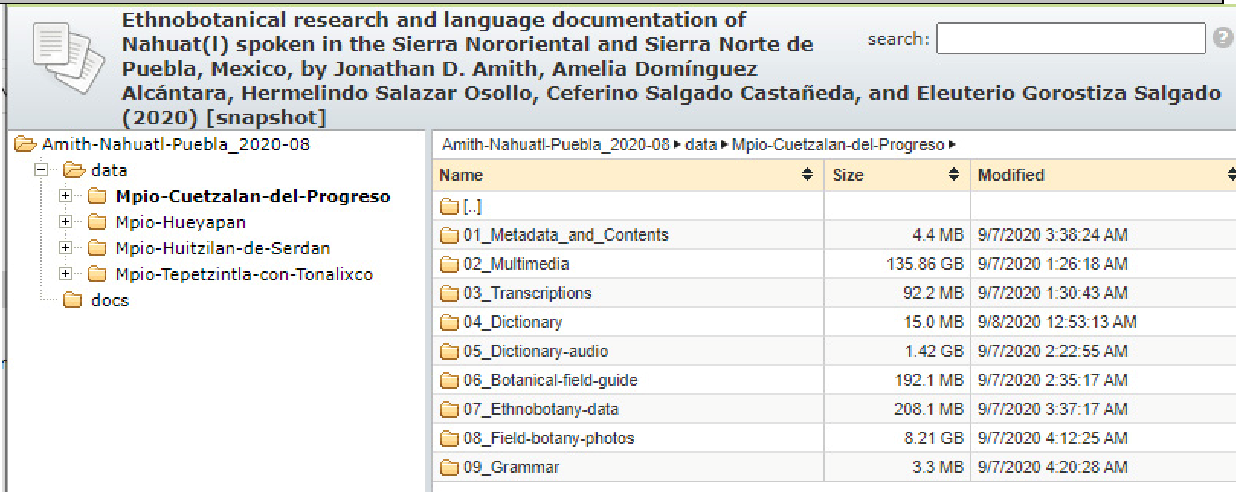
The general structure is that the data is divided into the four Nahuat[l]-speaking municipalities for which data was collected. The most extensive documentation was in the municipality of Cuetzalan del Progreso, for which 9 sub-folders were created (see image below). The remaining 3 municipalities have smaller corpora.

A **Snapshot** of the entire deposit is found in an HTML documented in the \docs folder entitled 11_Snapshot-of-Amith-Nahuat-deposit-file-structure-and-files.html

Following the Snapshot is a quick overview of the contents of each sub-folder. Note that there are also readme files in these subfolders that should be consulted as well

Directory Structure of Ethnobotanical research and language documentation of Nahuat(l) spoken in the Sierra Nororiental and Sierra Norte de Puebla

Overall deposit file structure showing \data and \docs.
Data is divided into 4 sub-folders, each with data pertaining a particular municipality



Ethnobotanical research and language documentation of Nahuat(l) spoken in the Sierra Nororiental and Sierra Norte de Puebla, Mexico, by Jonathan D. Amith, Amelia Domínguez Alcántara, Hermelindo Salazar Osollo, Ceferino Salgado Castañeda, and Eleuterio Gorostiza Salgado (2020) [snapshot]

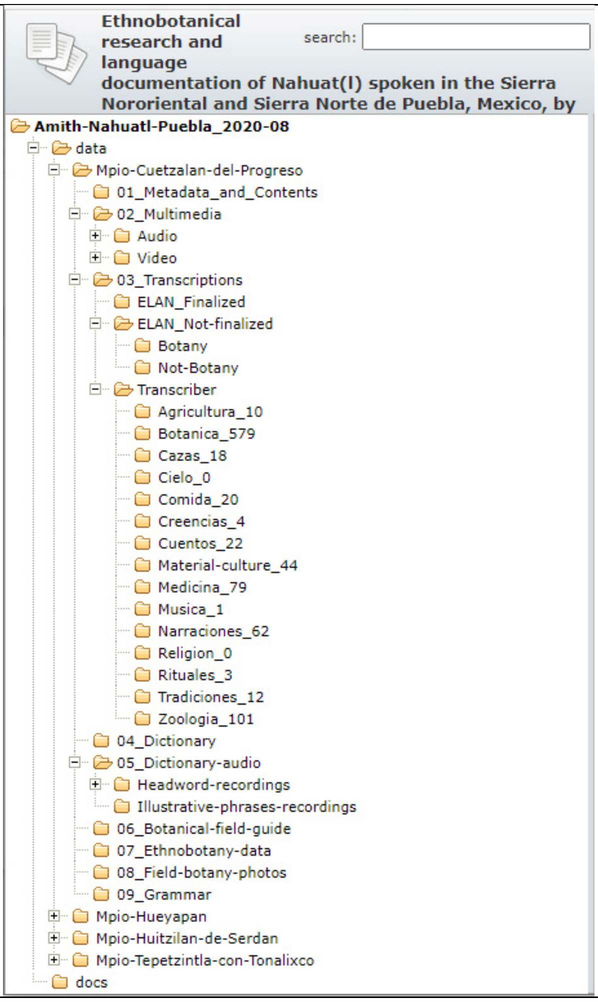
search:

- Amith-Nahuatl-Puebla_2020-08
  - data
    - **Mpio-Cuetzalan-del-Progreso**
    - Mpio-Hueyapan
    - Mpio-Huitzilan-de-Serdan
    - Mpio-Tepetzintla-con-Tonalixco
  - docs

Amith-Nahuatl-Puebla_2020-08 ▸ data ▸ Mpio-Cuetzalan-del-Progreso ▸

Name	Size	Modified
[..]		
01_Metadata_and_Contents	4.4 MB	9/7/2020 3:38:24 AM
02_Multimedia	135.86 GB	9/7/2020 1:26:18 AM
03_Transcriptions	92.2 MB	9/7/2020 1:30:43 AM
04_Dictionary	15.0 MB	9/8/2020 12:53:13 AM
05_Dictionary-audio	1.42 GB	9/7/2020 2:22:55 AM
06_Botanical-field-guide	192.1 MB	9/7/2020 2:35:17 AM
07_Ethnobotany-data	208.1 MB	9/7/2020 3:37:17 AM
08_Field-botany-photos	8.21 GB	9/7/2020 4:12:25 AM
09_Grammar	3.3 MB	9/7/2020 4:20:28 AM

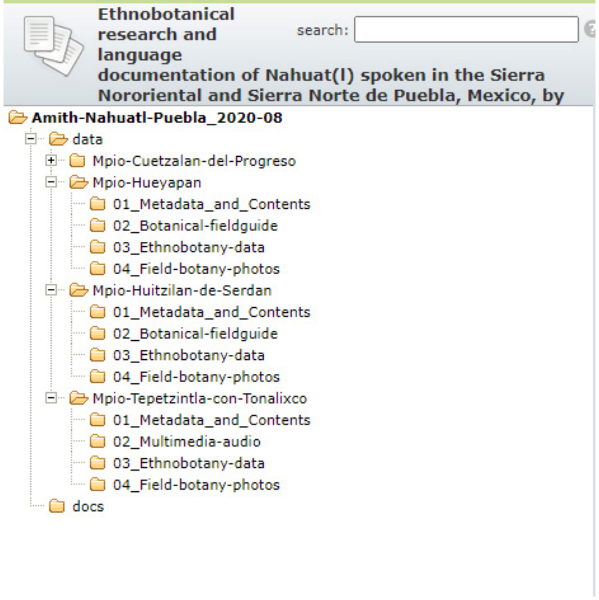
File structure of Municipality of Cuetzalan **9 distinct sub-folders based on theme/content**	**File structure of Remaining Municipalities** **4 distinct sub-folders based on theme/content**



Ethnobotanical research and language documentation of Nahuat(l) spoken in the Sierra Nororiental and Sierra Norte de Puebla, Mexico, by

search:

```
Amith-Nahuatl-Puebla_2020-08
  data
    Mpio-Cuetzalan-del-Progreso
      01_Metadata_and_Contents
      02_Multimedia
        Audio
        Video
      03_Transcriptions
        ELAN_Finalized
        ELAN_Not-finalized
          Botany
          Not-Botany
        Transcriber
          Agricultura_10
          Botanica_579
          Cazas_18
          Cielo_0
          Comida_20
          Creencias_4
          Cuentos_22
          Material-culture_44
          Medicina_79
          Musica_1
          Narraciones_62
          Religion_0
          Rituales_3
          Tradiciones_12
          Zoologia_101
      04_Dictionary
      05_Dictionary-audio
        Headword-recordings
        Illustrative-phrases-recordings
      06_Botanical-field-guide
      07_Ethnobotany-data
      08_Field-botany-photos
      09_Grammar
    Mpio-Hueyapan
    Mpio-Huitzilan-de-Serdan
    Mpio-Tepetzintla-con-Tonalixco
  docs
```



Ethnobotanical research and language documentation of Nahuat(l) spoken in the Sierra Nororiental and Sierra Norte de Puebla, Mexico, by

search:

```
Amith-Nahuatl-Puebla_2020-08
  data
    Mpio-Cuetzalan-del-Progreso
    Mpio-Hueyapan
      01_Metadata_and_Contents
      02_Botanical-fieldguide
      03_Ethnobotany-data
      04_Field-botany-photos
    Mpio-Huitzilan-de-Serdan
      01_Metadata_and_Contents
      02_Botanical-fieldguide
      03_Ethnobotany-data
      04_Field-botany-photos
    Mpio-Tepetzintla-con-Tonalixco
      01_Metadata_and_Contents
      02_Multimedia-audio
      03_Ethnobotany-data
      04_Field-botany-photos
  docs
```

\data

Mpio-Cuetzalan-del-Progreso

01_Metadata_and_Contents

This folder contains 3 sets of metadata for the multimedia files found in folder 02_Multimedia. The metadata files are csv and tab-delimited files (they have identical content). There are three basic "contents": (1) video; (2) audio with one consultant/speaker; (3) audio with two consultants/speakers. The audio has been divided by number of speakers to make sure the columns in the csv and tab-delimited files line up as a column is present for each consultant/speaker and an additional speaker means additional columns. The video is consistent in number of consultants

NOTE: Demographic data from all municipalities on the participants/consultants on this project is found in the \docs folder in the following files

03a_Puebla-Nahuatl-Totonac-consultant-database-for-deposit_2020-08-31.csv
03b_Puebla-Nahuatl-Totonac-consultant-database-for-deposit_2020-08-31_Tab-delimit.txt

Note that in the \docs folder there is a pdf catalogue of all the multimedia from Cuetzalan del Progreso. The file is named 08_Catalogue-of-multimedia-Sierra-Nororiental-de-Puebla.pdf

02_Multimedia

This folder contains one sub-folder for video and one for audio. The video sub-folder contains an additional 15 sub-folders, each representing a single video "shoot". The audio files are divided into 15 sub-folders based on themes that Amith uses in his documentation. Note that some sub-folders are emply (represented by _0 at end of folder name). This is because for this particular community/municipality no audio relevant to this these was recorded. Nevertheless, the folder structure is maintained.

03_Transcriptions

The transcriptions of the audio files is in process, which explains the distinct sub-folders under 03_Transcriptions. There are 954 speech recordings (with 1 additional music recordings). All 954 speech recordings have been initially transcribed in the program called Transcriber (see subfolder Transcriber). Of these 954 a total of 738 have been imported to ELAN. Of these 738 a total of 299 have been proofed and finalized in ELAN (folder ELAN_Finalized). The remaining 439 are still pending proofing by Amith (although the Transcriber transcriptions are very good). These 439 ELAN files are divided into Botany (279 files) and Not-Botany (160 files). See future plans for how these files will be finalized.

04_Dictionary

This is a draft of a Highland Puebla Nahuat dictionary in process. The two files (one txt and the other XML) have the same basic content. Note that the txt file was created in a program called Toolbox and then exported to XML. The file names are Active-Dictionary_2020-09-02.txt and Active-Dictionary_2020-09-02.xml.

The structure/fields of the dictionary are in flux but the following gives a general idea of the most important fields at present.

\lx HEADWORD a:a:wiltia
\lx_cita CITATION FORM (for recording) a:a:wiltia
\lx_var VARIANT (if limited to specific villages) General
\ref 04126 Unique identifier
\glosa STANDARDIZED GLOSS jugar.en.agua
\catgr Grammatical category V1
\deriv DERIVATIONAL PATTERN
\infl INFLECTIONAL CLASE (Verbs and nouns, in particular) Clase 2a
\diag DIAGNOSTIC, i.e., whether stem takes these prefixes -na:l-; +sen-; -pan-; -ye:k-; -tel-; -ta-; -lo
\sig MEANING jugar con o en el agua
\osten OSTENTIVE DEFINITION p. ej., al meterse en un charco o un río, al agarrar una cubeta llena de agua para divertirse
\sig_var VARIANT OF MEANING General
\fr_n NAHUAT ILLUSTRATIVE PHRASE Nopili a:a:wilitih wa:n a:paliw. Yehwa ika niktapatilih.
\fr_au AUTHOR OF ILLUSTRATIVE PHRASE (speaker code) EGS301
\fr_var VARIANT OF ILLUSTRATIVE PHRASE Tzina
\fr_e SPANISH OF ILLUSTRATIVE PHRASE Mi hijo jugó con agua y se mojó. Por eso le cambié la ropa.
\fr_son SOUND FILENAME OF ILLUSTRATIVE PHRASE 04126-01_EGS-Tzina
\fr_fuente SOURCE OF ILLUSTRATIVE PHRASE (from corpus or created/edited) Frase elaborada
\fr_n Nopili a:mo takaki. Ihwa:k nitapa:kati a:taw a:a:wiltia wa:n moa:palowa.
\fr_au EGS301
\fr_var Tzina
\fr_e Mi hijo es inquieto. Cuando voy a lavar ropa al río juega con el agua (en el agua) y se moja.
\sig (con <nawa>sen-</nawa> : <nawa>sena:a:wiltiah</nawa> | jugar con o en el agua juntos, en grupo
\fr_n Mah mitstapa:ki mokni:w. Ta: nansena:a:wilti:toh ya:lwa, ompa ximotapa:kia:nih ya.
\fr_au ADA300
\fr_var Xaltn
\fr_e ¡Que tu hermano lave tu ropa! Así como ayer fueron juntos a jugar en el agua (en el rio), ahi hubieran ya lavado tu ropa.
\fr_son Pendiente grabar
\fr_fuente Frase elaborada
\raiz ROOT a:
\raiz ROOT a:wil
\nsem NOTES ON SEMANTICS
\nmorf NOTES ON MORPHOLOGY La raiz <root>a:</root> es un objeto incorporado, reduciendo la valencia del verbo transitivo <vnawa>a:wiltia</vnawa>
\ncomp COMPARATIVE NOTES

\dt DATE LAST EDITED 12/Feb/2020

05_Dictionary-audio

Eventually the dictionary will have two basic types of associated files: headword recordings (repeated 3 times per file) and illustrative sentences. Initial material is in this deposit in separate folders. Note that the illustrative sentences are of two types: (a) from the corpus (found through concordancing); (b) elaborated/edited by native speakers and Amith to illustrate specific senses. Further details are found in the file ReadMe_Dictionary-audio_Cuetzalan.txt

06_Botanical-field-guide

This is simply a pdf file of all plants collected in the municipality with scientific names, Indigenous names (when they exist), comparative data on names, plant uses, and eventually descriptions of the plants. These guides are in continual revision and will be updated in subsquent editions of this corpus. Each municipality has one guide.

07_Ethnobotany-data

Ethnobotanical data comprises strictly botanical data (the collection data that botanists include in their notes, databases, and labels) and ethnobotanical (the nomenclature, classification, and use of plants, reported for each consultant present at the time of collection). These two sets of data are included in this folder in the following ways

1. Label information: these are labels (pdf) for the entire collection for this municipality, There are two files, one sorted by collection number and one sorted by Family and species
   Sierra-Nororiental-de-Puebla-Ethno-botany_Labels-sorted-by-collection-number_2020-08-31.pdf
   Sierra-Nororiental-de-Puebla-Ethno-botany_Labels-sorted-by-Family-species_2020-08-31.pdf
2. Complete data in txt files. These are two text files (one csv and the other tab-delimited) that contain all the botanical and ethnobotanical data from the database. The two files are
   Sierra-Nororiental-de-Puebla-ethno-botany-research_Complete-data_2020-08-31.csv
   Sierra-Nororiental-de-Puebla-ethno-botany-research_Complete-data_Tab-delim_2020-08-31.txt
3. Subset of the complete data. This subset contains (by consultant) the nomenclature, classification, and use of plants, with particular attention to plant name analysis (morphemes) and gloss (meaning of each morpheme). Again, data is presented in different formats, in this case 3
   Sierra-Nororiental-de-Puebla-ethno-botany-research_Ethnographic-data_2020-08-31.csv
   Sierra-Nororiental-de-Puebla-ethno-botany-research_Ethnographic-data_Tab-delim_2020-08-31.txt
   Sierra-Nororiental-de-Puebla-ethno-botany-research_Ethnographic-data_2020-08-31.mht
4. Summary presentation of collection number: This file is particularly important for the following folder of plant photos in situ. Plant photo filenames include the collection number as a reference. This summary file lists only the colleciton number, family, and species for quick reference / look-up of photos
   Sierra-Nororiental-de-Puebla-ethno-botany-research_Collection-number-Scientific-names_Tab-delim_2020-08-31.txt

08_Field-botany-photos

The files in this folder are photographs (3,767). The semantics of the file name is explained in the Readme file. Note that for a quick identification of the species photographed use the summary presentation in the previous folder: Sierra-Nororiental-de-Puebla-ethno-botany-research_Collection-number-Scientific-names_Tab-delim_2020-08-31.txt See Readme file in this folder for further details.

09_Grammar

These are 13 pdf files that are drafts (mostly in Spanish at this point) of a grammar in process. As time passes these chapters will be edited and translated into English.

Mpio-Hueyapan
Mpio-Huitzilan-de-Serdan

These two sub-folders are included together as their content is virtually identical in general structure. Each has four subfolders and the content of each of these sub-folders mirrors the content of parallel folders in the Mpio-Cuetzalan-del-Progreso. See also the Readme files in each subfolder for more detail.

01_Metadata_and_Contents

The contents of these sub-folders is simply a Readme file as to date there are no multimedia recordings from these two municipalities.

02_Botanical-fieldguide

The contents these sub-folders are very similar to that of the equivalent sub-folder in regard to Cuetzalan del Progreso (named 06_Botanical-field-guide)

03_Ethnobotany-data

The contents these sub-folders are very similar to that of the equivalent sub-folder in regard to Cuetzalan del Progreso (named 07_Ethnobotany-data)

04_Field-botany-photos

The contents these sub-folders are very similar to that of the equivalent sub-folder in regard to Cuetzalan del Progreso (named 08_Field-botany-photos)

Mpio-Tepetzintla-con-Tonalixco

These two sub-folders are included together as their content is virtually identical in general structure. Each has four subfolders and the content of each of these sub-folders mirrors the content of parallel folders in the Mpio-Cuetzalan-del-Progreso. See also the Readme files in each subfolder for more detail.

01_Metadata_and_Contents

The contents of these sub-folders are very similar to that of the equivalent sub-folder in regard to Cuetzalan del Progreso (also named 01_Metatadata_and_Contents)
Note that in the \docs folder there is a pdf catalogue of all the multimedia from Tepetzintla. The file is named 09_Catalogue-of-multimedia-Nahuatl-in-Municipality-of-Tepetzintla.pdf

02_Multimedia-audio

The major difference of these 151 files is that they are all botanical field recordings, i.e,. field recordings taken at the time of plant collections. Each consultant (in this case a woman and her daughter) were recorded one after the other. For the metadata of these recordings and the summary pdf catalogue of selected metada, see the description of the sub-folder 01_Metadata_and_Contents

03_Ethnobotany-data

The contents this sub-folder is very similar to that of the equivalent sub-folder in regard to Cuetzalan del Progreso (named 07_Ethnobotany-data)

04_Field-botany-photos

The contents this sub-folder is very similar to that of the equivalent sub-folder in regard to Cuetzalan del Progreso (named 08_Field-botany-photos)

\docs

There are 13 files in the docs folder. The following is a brief description of each.

01_Deposit-overview_2020-08-31.pdf

This is this present document.

02_CV_JAmith active.pdf

Self-explanatory: Most recent Amith CV.

03a_Puebla-Nahuatl-Totonac-consultant-database-for-deposit_2020-08-31.csv

03b_Puebla-Nahuatl-Totonac-consultant-database-for-deposit_2020-08-31_Tab-delimit.txt

All transcriptions and audio filenames include reference to the consultants who were recorded. The consultants are identified by a 6-character code (initials + progressive number). The consultant demographic data (village of origin, birth year) are given in these two files.

04_File-naming-codification_2020-08-31.pdf

Description of the semantics of filenames for audio and transcriptions, in particular.

05_Transcription-conventions-J-Amith-2020-08-31.pdf

Brief notes on the guide to transcription (e.g., use of ellipsis).

06_Botany-Ethnobotany-metadata-fields-and-contents_2020-08-31.pdf

This is a field-by-field (75 fields in total) description of the metadata relevant to botanical collections.

07_Multimedia-metadata-fields-and-contents_2020-08-31.pdf

This is a field-by-field (37 fields in total) description of the metadata relevant to recordings and transcriptions/translations.

08_Catalogue-of-multimedia-Sierra-Nororiental-de-Puebla.pdf

This is a pdf catalogue for the municipality of Cuetzalan del Progreso of some of the most important descriptive information in the metadata for recordings and transcriptions/translations.

09_Catalogue-of-multimedia-Nahuatl-in-Municipality-of-Tepetzintla.pdf

This is a pdf catalogue for the municipality of Tepetzintla of some of the most important descriptive information in the metadata for recordings and transcriptions/translations.

10_EAF_Annotation_Format_3.0_and_ELAN.pdf

Many of the transcriptions/translations are in ELAN format (file extension is .eaf). This pdf describes the structure of the XML .eaf file.

11_Snapshot-of-Amith-Nahuat-deposit-file-structure-and-files.html

This is an HTML document that captures the folder and file structure of the deposit. The snapshot offers a for each folder a view of every file in the folder. For each file it displays the filename and size.

EAFv3.0.xsd

This is the schema definition for the ELAN (.eaf) files that are included in this deposit for the transcription and translation (Spanish) of the Nahuat[l] corpus

trans-14.dtd

This is the document type definition file for the Nahuat[l] transcriptions that are in Transcriber format.

7. Filename conventions

The conventions for naming audio and transcription files is presented in 04_File-naming-codification_2020-08-31.pdf. For other types of files (e.g., the in situ botanical photos) the filenaming conventions may be presented in Readme files in the relevant folder. No prohibited characters are included in the filenames.

8. Data formats

Text files: UTF-8. With one set of exceptions, all text files are in UTF-8. Most transcriptions (and the most up-to-date transcriptions) are in ELAN (.eaf) XML format.

ISO-8859-1. The 969 Transcriber files (which have the extension .trs) are all in ISO 8859-1. A list of these files is found in the file named Transcribir-trs-files-in-ISO-8859-1.txt located in the \docs folder. Note that even though the .dtd for Transcriber (trans-14.dtd in the \docs folder) says <?xml encoding="UTF-8"?> the files open up properly in Transcriber with the ISO 8859-1 encoding but not in UTF-8.

Audio: The vast majority of recordings are 16bit, 48KHz sampling rate. They are in uncompressed .wav format. A few files may have been (erroneously) recorded at 44.1 KHz sampling rate.

The Zoom H4n field recordings (made at the time a plant was collected with a Zoom H4n; all filenames include the code "BotFl" for the "genre" in the filename. Note that these recordings are single channel made with the Zoom internal microphone. A few files might be at 44.1KHz. The metadata notes this when it occurs. Please also note that due to the fact that the recordings were made by native speakers who were using this recorder for the first time, there was some clipping (files were normalized, usually to 67%). Although the field recordings have not been transcribed they have been described in detail. These descriptions (which were only made in Tepetzintla) are found in the 01_Metadata-and-Contents folder (see Metadata_Tepetzintla_Audio-recordings-with-2-consultants_2020-08-31). See also the same descriptions in the relevant pdf catalogue: 09_Catalogue-of-multimedia-Nahuatl-in-Municipality-of-Tepetzintla.pdf

Marantz PMD 671 and Sound Devices 722 recordings: These recordings are generally conversations among two speakers. In these cases each speaker is recorded with a separate microphone (head- or earworn) and on a distinct channel. A few recordings (e.g., fictional stories) might have one speaker and one channel. Again the format is 16-bit and 48KHz sampling. The metadata gives information on which speaker is on which channel.

Video: All video files are mp4.

PDF: PDF files, in the \docs folder, are all pdf/a

XML: There are three sources of XML files.

The first is the program ELAN whose files have the extension .eaf. The XSD associated with this files is found in EAFv3.0.xsd See also the exaplanation of the format in the file 10_EAF_Annotation_Format_3.0_and_ELAN.pdf Both files are in this \docs folder

The second source of XML files is that found in Transcriber files (.trs) the DTD is included in the \docs folder and is named trans-14.dtd

The third source of XML files is the export of a Toolbox txt file. There is no DTD for the dictionary XML file (in the \data\Mpio-Cuetzalan-del-Progreso\04_Dictionary folder, Active-Dictionary_2020-09-02.xml

9. Automatic Speech Recognition corpus and recipe location

Portions of this project have been used (starting in 2020) to develop end-to-end automatic speech recognition of Highland Puebla Nahuatl.

The corpus as it was used for this effort has been deposited/uploaded at https://www.openslr.org/92/

The ASR recipe is found at https://github.com/espnet/espnet/tree/master/egs/puebla_nahuatl/asr1