Supplementary Material 2: Orthography conversion, cognate coding, Q-residuals and Delta scores

Subgrouping in a 'dialect continuum': A Bayesian phylogenetic study of the Mixtecan language family. Journal of Language Evolution

Sandra Auderset, Simon J. Greenhill, Christian T. DiCanio, Eric W. Campbell

February 15, 2023

Contents

1 Language names and codes

The branch name 'Mixtec' is often used as a cover term for all Mixtec varieties/peoples as well as for individual varieties/groups and is a Nahuatl exonym roughly translating to 'inhabitants of the cloudy place' (from *miš-* 'cloud' and *te·ka* 'inhabitant of place of', cf. Campbell 1997:402). Mixtec people themselves use it with non-Mixtec people and it does not seem to be perceived as derogatory or negative. Mixtec endonyms, which are rarely given in publications, are often near identical or very similar to each other. Although no systematic information is available, many Mixtec groups refer to themselves as 'people of the rain' (*ñuu savi*) and to their language as 'words/language of the rain' (*tu'un/sa'an savi/sau*), but 'humble words/language' (*tu'un nta'vi*) is also common. The latter is viewed as problematic by some Mixtec language activists, especially when used by outsiders, since the Mixtec expression could also be translated as 'poor words/language'. We thus use the exonyms, but always together with the full village name of where the variety is spoken. The practice of omitting the patron saint part of the village name might seem practical for saving space, but creates confusion in cases like San Miguel Peras and San Martín Peras or Magdalena Peñasco, San Mateo Peñasco, and Guadalupe Peñasco etc. (cf. also Okura 2015). A similar situation is found in Cuicatec and Triqui, although there is less confusion between varieties simply because there are fewer (two for Cuicatec and three for Triqui according to Hammarström et al. 2021). Cuicatec is also a Nahuatl exonym and roughly translates to 'inhabitants of the place of song' (from *kwi·ka* 'song' and *te·ka* 'inhabitant of place of', cf. Campbell 1997:402). For Triqui (alternatively spelled Trique or Triki) no information on the origin of this denomination is available.

The situation cannot be resolved by using Glottocodes (Hammarström et al. 2021) or ISO-codes (Eberhard et al. 2021). These codes often bundle varieties together which are spoken in the same municipality or region without being explicit about which villages the code includes. Furthermore, our study includes various varieties for which neither system has assigned a code. As mentioned above, we thus use the full village name plus branch name in our study, but for convenience, I also developed a system of codes that uniquely identifies each Mixtec-, Triqui-, or Cuicatec-speaking village. The codes consist of the first four letters of the village name excluding the patron saint and two numbers. All contemporary varieties have a number that starts with 1, while historical varieties start with 0. The second number is meaningless and only serves to disambiguate varieties for which the letters and first number coincide. An overview of the sample languages with the codes, as well as Glottcodes and ISO-codes if available is provided below in Table 1 and available for download as a tsv.

Table 1: Language names and codes

Code	Village Name	Branch	Glottocode	ISOcode
abas11	Abasolo del Valle	Mixtec	mixt1425	mix
acat11	Santa María Acatepec	Mixtec	sant1441	mtu
achi11	San Miguel Achiutla	Mixtec	sanm1292	xtm
adeq11	Santa Catarina Adéquez	Mixtec		
ahue11	San Miguel Ahuehuetitlán	Mixtec	sila1250	mks
alac11	Alacatlatzala	Mixtec	alac1244	mim
alco11	Alcozáuca de Guerrero	Mixtec	alco1235	xta
amat11	San Miguel Amatitlán	Mixtec		
amol11	Santiago Amoltepec	Mixtec	amol1236	mbz
amol12	San Cristóbal Amoltepec	Mixtec	sanc1244	xtm

Code	Village Name	Branch	Glottocode	ISOcode
apas11	Santa María Apazco/Apasco	Mixtec	apas1235	mip
apoa11	Santiago Apoala	Mixtec	apas1235	mip
asun11	Santa María Asunción	Mixtec		
atat11	San Esteban Atatlahuca	Mixtec	atat1238	mib
aten11	San Agustín Atenango	Mixtec		
atla11	Atlamajalcingo del Monte	Mixtec	atla1271	mim
atoy11	San Pedro Atoyac	Mixtec		
ayut11	Ayutla de los Libres	Mixtec	ayut1236	miy
bate11	La Batea	Mixtec	mixt1425	mix
brav11	Santa Cruz de Bravo	Mixtec		
caca11	Santiago Cacaloxtepec	Mixtec	caca1250	miu
cahu11	Cahuatache/Cahuatachi	Mixtec	alac1244	mim
chal11	Chalcatongo de Hidalgo	Mixtec	sanm1296	mig
chay11	San Agustín Chayuco	Mixtec	chay1249	mih
chay12	San Pedro Chayuco	Mixtec		
chaz11	Santiago Chazumba	Mixtec	chaz1235	xtb
chic02	San Andrés Chicahuaxtla 1890	Triqui	chic1273	trs
chic11	San Miguel Chicahua	Mixtec		
chic12	San Andrés Chicahuaxtla	Triqui	chic1273	trs
chig11	Chigmecatitlán	Mixtec	chig1239	mii
coat11	San Juan Coatzóspam/Coatzóspan	Mixtec	coat1241	miz
coic11	Coicoyán de las Flores	Mixtec	coic1238	jmx
colo11	San Juan Colorado	Mixtec	sanj1281	mjc
conc11	Concepción Pápalo	Cuicatec	tepe1280	cux
copa11	San Juan Copala	Triqui	copa1237	trc
coso11	Cosoltepec	Mixtec	chaz1235	xtb
coxc11	San Pedro Coxcaltepec Cántaros	Mixtec		
cris11	San Cristóbal (Jamiltepec)	Mixtec	chay1249	mih
cuat11	Cuatzoquitengo	Mixtec	cuat1239	mim
cuau11	Santa Ana Cuauhtémoc	Mixtec	cuya1240	xtu
cuil11	Cuilapam/Cuilapan de Guerrero	Mixtec		
cuya11	Cuyamecalco Villa de Zaragoza	Mixtec	cuya1240	xtu
diux11	San Juan Diuxi	Mixtec	diux1236	xtd
dura11	San Martín Duraznos	Mixtec	juxt1235	vmc
esta11	San Martín del Estado	Mixtec		
este11	Santa Catarina Estetla	Mixtec	peno1244	mil
flor11	San Francisco de las Flores	Mixtec		
flor12	San Marcos de la Flor	Mixtec	sanm1291	jmx
gran11	San Miguel El Grande	Mixtec	sanm1295	mig
gran12	San Miguel El Grande (Alcozáuca)	Mixtec	alco1235	xta
guad11	Guadalupe Villahermosa (El Portesuelo)/Guadalupe Portezuelo	Mixtec		
huaj11	Heroica Ciudad de Huajuapan de León	Mixtec	caca1250	miu
huaz11	Santa María Huazolotitlán	Mixtec		
huen11	Santo Domingo Huendío	Mixtec	sant1440	xtm
huit11	San Antonio Huitepec	Mixtec	huit1252	mxs

Code	Village Name	Branch	Glottocode	ISOcode
itun11	Santa Cruz Itundujia	Mixtec	itun1239	mce
itun12	San Martín Itunyoso	Triqui	sanm1298	trq
ixpa11	Ixpantepec Nieves	Mixtec	sila1250	mks
ixta11	Santiago Ixtayutla	Mixtec	ixta1235	vmj
ixta12	Santiago Ixtaltepec	Mixtec		
ixta13	San Miguel Ixtapam/Ixtapan	Mixtec		
jami11	Santiago Jamiltepec	Mixtec	jami1235	mxt
jate11	San Francisco Jaltepetongo	Mixtec		
jica11	Santa María Jicaltepec	Mixtec	pino1237	mio
jica12	San Pedro Jicayán	Mixtec		
jica13	El Jicaral	Mixtec	coic1238	jmx
joco11	San Pedro Jocotipac	Mixtec		
juxt11	Santiago Juxtlahuaca	Mixtec	juxt1235	vmc
lore11	San Lorenzo	Mixtec		
mech11	Santa Catarina Mechoacán	Mixtec	chay1249	mih
metl11	Metlatónoc	Mixtec	metl1238	mxv
mica11	El Rosario Micaltepec	Mixtec		
mitl11	Santiago Mitlatongo	Mixtec	mitl1235	vmm
mixt11	San Juan Mixtepec	Mixtec	mixt1425	mix
moli11	San Pedro Molinos	Mixtec	sanp1259	mig
mont11	San Antonino Monte Verde	Mixtec	mont1271	xtn
mont12	San Sebastián del Monte	Mixtec	sila1250	mks
mont13	Santa Lucía Monteverde	Mixtec	sant1439	mdv
mont14	San Agustín Monte Lobos	Mixtec		
more11	San Luís Morelia	Mixtec	tama1339	vmx
ndua11	San Antonio Nduayaco	Mixtec	apas1235	mip
nica11	San Sebastián Nicananduta	Mixtec	nort2985	xtn
nuch11	San Jorge Nuchita	Mixtec	sila1250	mks
numi11	San Juan Ñumi	Mixtec	nort2985	xtn
nund11	Santa Cruz Nundaco	Mixtec		
nund12	Santiago Nundiche/Nundichi	Mixtec	nort2985	xtn
nund13	Santo Domingo Nundó	Mixtec		
nuti11	Santa María Nutio	Mixtec		
nuxa11	Santo Domingo Nuxaá	Mixtec		
nuxi11	San Andrés Nuxiño	Mixtec		
nuyo11	Santiago Nuyoó	Mixtec	nuyo1238	meh
ocot11	Santo Tomás Ocotepec	Mixtec	ocot1243	mie
papa11	Santa María Pápalo	Cuicatec	sant1437	cux
pena11	Magdalena Peñasco	Mixtec	magd1235	xtm
pena12	San Mateo Peñasco	Mixtec	sanm1293	xtm
peno11	Santa María Peñoles	Mixtec	peno1244	mil
peno11	Santa María Peñoles	Mixtec	sant1445	mil
pera11	San Martín Peras	Mixtec	sanm1291	jmx
petl11	Petlacalancingo/Petlakalansingo	Mixtec	petl1238	xta
pied11	San Miguel Piedras	Mixtec	sanm1294	xtp
pied12	Piedra Azul	Mixtec		
pina11	San Juan Piñas	Mixtec	sanj1280	jmx

Code	Village Name	Branch	Glottocode	ISOcode
pino11	Santiago Pinotepa Nacional	Mixtec	pino1237	mio
pino12	Pinotepa de Don Luís	Mixtec	pino1237	mio
plan11	Plan de Guadalupe	Mixtec	plan1238	mim
poto11	Potoichán	Mixtec	poto1253	mim
prog11	San Miguel Progreso	Mixtec		
prog12	San Jerónimo Progreso	Mixtec	sila1250	mks
prom01	Proto-Mixtec	Mixtec	mixt1427	
prot01	Proto-Trique	Triqui	triq1251	
reye11	Santos Reyes Tepejillo	Mixtec		
sayu11	San Francisco (de Asís) Sayultepec	Mixtec	pino1237	mio
sila11	Silacayoapam	Mixtec	sila1250	mks
sind11	San Mateo Sindihui	Mixtec	sind1277	xts
sini11	San José (Sinicahua)	Mixtec	sini1243	xti
soso11	San Jerónimo Sosola	Mixtec		
soya11	San Bartolo Soyaltepec	Mixtec	soya1236	vmq
tama11	San Juan Tamazola	Mixtec	tama1339	vmx
tama12	Santiago Tamazola	Mixtec		
tata11	Santa María Tataltepec	Mixtec	sant1444	
teco11	San Sebastián Tecomaxtlahuaca	Mixtec		
teit11	San Juan Teita	Mixtec	sanj1282	xtj
tejo11	Los Tejocotes	Mixtec		
tepa11	Tepango	Mixtec	ayut1236	miy
tepa12	San Mateo Tepantepec	Mixtec	sanm1297	mil
tepe03	San Juan Tepeuxila 1900	Cuicatec	tepe1280	cux
tepe11	Tepejillo	Mixtec		
tepe12	San Antonio Tepetlapa	Mixtec		
tepe13	San Juan Tepeuxila	Cuicatec	tepe1280	cux
tepo01	San Pedro y San Pablo Teposcolula 1600	Mixtec		
teut11	San Pedro Teutila	Cuicatec	teut1235	cut
tezo11	Tezoatlán de Segura y Luna	Mixtec	tezo1239	mxb
ticu11	Santa Catarina Ticuá	Mixtec	sant1443	mig
tida11	San Pedro Tidaá	Mixtec	tida1235	mtx
tija11	San Pablo Tijaltepec	Mixtec	tija1235	xtl
tila11	Santiago Tilantongo	Mixtec	diux1236	xtd
tind11	Santa María Tindú	Mixtec		
tlac11	San Agustín Tlacotepec	Mixtec	sana1284	xtm
tlac12	San Miguel Tlacotepec	Mixtec		
tlah11	Tlahuapa	Mixtec	alco1235	xta
tlal11	Santa Catarina Tlaltempan	Mixtec		
tlaz11	Santiago Tlazoyaltepec	Mixtec	tlaz1235	mqh
tlaz11	Santiago Tlazoyaltepec	Mixtec	sant1446	mqh
tona11	Santo Domingo Tonahuixtla	Mixtec		
toto11	Tototepec	Mixtec	toto1307	mim
toto12	(Santa María) Totoltepec de Guerrero	Mixtec	chaz1235	xtb

Code	Village Name	Branch	Glottocode	ISOcode
tutu11	Villa de Tututepec de Melchor Ocampo	Mixtec	tutu1243	mtu
tutu12	San Pedro Tututepec	Mixtec	tutu1243	mtu
xaya11	San Jerónimo Xayacatlán	Mixtec	south3001	mit
xaya12	Xayacatlán de Bravo	Mixtec	south3001	mit
xoch11	Xochapa	Mixtec	xoch1238	xta
yolo11	Santa María Yolotepec	Mixtec	taca1257	xtt
yolo12	Yoloxochitl	Mixtec	yolo1241	xty
yoso11	Santiago Yosondúa	Mixtec	yoso1239	mpm
yoso12	Santa María Yosoyúa	Mixtec	sant1442	
yoso13	Santa Catarina Yosonotú	Mixtec	sant1438	
yoso14	San Pedro Yosoñama	Mixtec	yoso1238	xtn
yucu11	San Bartolomé Yucuañe	Mixtec	yucu1250	mvg
yucu12	Santa María Yucuhiti	Mixtec	yucu1249	meh
yucu13	Yucuñuti de Benito Juárez	Mixtec	yucu1251	mxb
yucu14	San Pedro Yucunama	Mixtec		
yucu15	Santa María Yucunicoco	Mixtec		
yucu16	Yucuquimi de Ocampo	Mixtec		
yucu17	Yucunani/Yucunany	Mixtec	mixt1425	mix
yuta11	Yutanduchi de Guerrero	Mixtec	yuta1240	mab
yuta12	San Juan Yuta	Mixtec		
yuta13	San Andrés Yutatío	Mixtec		
zaca11	Santa María Zacatepec	Mixtec	sant1436	mza
zahu11	San Simón Zahuatlán	Mixtec	sans1274	
zapo11	Zapotitlán Palmas	Mixtec		

2 Converting orthography to IPA

This section contains an overview of how the orthographies of each source were converted to (a standardized version of) the International Phonetics Alphabet (IPA). This step is necessary for adequately comparing and aligning entries to identify cognate sets (and later on sound changes). The orthographies used in materials on Mixtecan languages are highly diverse. Often each author and each source uses a system differing from all others in certain aspects. Nevertheless, one can identify three major groups of orthographical systems:

1. Orthographies based on Spanish: This set of orthographies tries to stay as close to the Spanish system as possible. The sound [k], for example, is represented by <c> or <qu> depending on the following vowel. Additions are made as necessary, usually with digraphs. Nasality on vowels, for example, is represented by an <n> following the vowel in question. The glottal stop is usually written as <h>. This system is most often found in pedagogical materials and/or materials developed by SIL missionaries.
2. Orthographies based on the Americanist Phonetic Notation: The Americanist Phonetic Notation, also called APA, is a phonetic system originally developed for the transcription of Native American languages (cf. the summary by the Western Institute for Endangered Language Documentation). It relies heavily on diacritics and mixes Latin and Greek characters. The sound [ʃ], for example, is represented as <š>. The glottal stop is usually written with the IPA symbol.This system is found in older sources and those predominantly geared to an academic audience.
3. Orthographies based on neither of the above: Some newer documentation efforts use a system that could be referred to as latin-based, i.e. it uses Latin characters with few diacritics but does not emulate Spanish orthography. The glottal stop is usually represented by an apostrophe or saltillo. Nasality on vowels is usually represented by an <n> following the vowel in question.

In all of these orthographic systems, some graphemes are easy to convert to IPA because there is no ambiguity as to what sound they represent. This is the case, for example, for the nasals *n*, *m*, and *ñ* and for most vowels. Other graphemes, however, are notoriously ambiguous also due to differences in the sound systems of these languages. These are:

- <y which sometimes represents [j], but other times [ʒ]
- <x> which often represents [ʃ], but is other times used for [x] or [ɕ]
- <h> which is used as the glottal stop in Spanish-based orthographies, but in others often represents [h]
- <j> which is sometimes used as [h] other times as [j]

Below I elaborate on some general principles for conversion and then address the specifics of each doculect beyond those general principles. In the detailed descriptions, I will generally only comment on the ambiguous graphemes. The doculects are ordered first by branch (Cuicatec, Triqui, Mixtec) and within that alphabetically by language code. For practical reasons, graphemes will not be written within angle brackets but presented in italics. IPA correspondences are rendered in square brackets as is customary (e.g. the grapheme *i* represents IPA [i]).

2.1 General Conversion Principles

2.1.1 Prenasalized stops vs. nasal + stop sequences

Mixtec and Cuicatec stops do not exhibit a voicing distinction, but are often described as opposing a series of voiceless stops to a series of prenasalized stops. However, some scholars have analyzed these as combinations of nasal and stop as sequences, rather than a single phoneme (so for example Pike & Oram (1976) and Macaulay (1996)). On empirical grounds, this issue can only be resolved by a detailed analysis of the distributional evidence in each variety, which lies outside the scope of this work. For practical purposes, i.e. for sequence alignments, consistency is the most important consideration, but there are no obvious advantages to representing all as prenasalized stops or nasal + stop sequences. I have decided to adopt the former solution, because it simplifies syllable structure, which could be helpful for automatic processing of the data. In addition, prenasalized stops are represented with voiced IPA symbols, even though they might not be fully voiced. Since there is no voicing opposition in stops, this is simply a matter of practicality and consistence.

2.1.2 Long vowels

Sequences of identical vowels are represented as such and not as a long vowel, e.g. aa is represented as [aa] and not [a:]. I do this for two reasons: 1) The tone bearing unit in Mixtecan is the mora, so that it is easier and less confusing to represent tone with each vowel written separately, and 2) There is variation between and within Mixtec varieties in words of the form $CV_iʔV_i$, such that these can contract to CV_iV_i. For comparative purposes, it is more straightforward to align across such variation if the vowels are not written as one segment.

2.1.3 Fortis/lenis opposition in Triqui consonants

In most Triqui varieties consonants show a fortis-lenis opposition. This opposition is best retained in Itunyoso Triqui, which shows a clear and consistent distinction between geminate and non-geminate consonants. The fortis-lenis constrasts in other varieties – as far as they are present – have been represented in the same way, i.e. as geminate vs. non-geminate consonants.

2.1.4 Glottal stop and glottalization

In Cuicatec and Triqui – and to a lesser extent in some Mixtec varieties – the glottal stop has a different status depending on its position and/or phonological processes it participates in. Usually, there is an opposition between the glottal stop as a consonant phoneme and glottalized vowels and consonants. The former is restricted to intervocalic position, while the latter can appear as pre- or post-glottalization. While this distinction could be important for establishing sound correspondences, it is less so for establishing cognates. For ease of of conversion this difference has been ignored and all instances are converted to [ʔ]. In addition, CLTS does not currently allow preglottalized consonants.

2.1.5 Tone representations

The IPA offers two principal ways of displaying tone: diacritics that are placed above the tone bearing unit (TBU) or tone bars, with the latter method suggested as the preferred one. Tone diacritics are useful in practical orthography but not well suited to alignments, for which it is

Table 2: Overview of tone notation based on Chao (1930)

Chao's number	Musical comparison	Label
5	G#	high
4	F	half-high
3	E	mid
2	D	half-low
1	C	low

more useful to represent the tone after the TBU as its own character (even if that is not the phonetic reality). Tone bars are difficult to type and read and thus not well suited for the task at hand. I use Chao's tone numbers (Chao 1930) instead, since they are widely known and easy to type and read.

In this system, each distinctive pitch level is assigned a number from one to five, with one being the lowest and five the highest, cf. Table 2. The interval between the lowest and highest pitch is assumed to correspond roughly to an augmented fifth. Contour tones are represented as combinations of these levels. A high to low falling tone, for example, would be noted as 51, while a low to mid rising tone would be noted as 13. This is not meant as a statement regarding their status as tonemes, but rather as a practical decision facilitating comparison between languages. The reader should keep this in mind when consulting the profiles, since this means that the profile does not necessarily tell you how many tonemes a language has.

In the source materials, we find a multitude of tone notations (if tones are noted at all). Most descriptions of varieties with three tonemes use diacritics, while those with more tonemes are usually represented with numbers. Below I summarize the most common systems.

Diacritics: Notations with diacritics are most common in languages with three level tonemes and five tonemes of which two are contour tones. But they also occur with four level tonemes. High tone is invariable marked with an acute accent, rising tone with a caron, and falling tone with a circumflex. The marking of mid and low tones varies across sources and usually one of them is left unmarked. If the mid tone is marked, this is usually done with a macron. If the low tone is marked a grave accent, underbar, or more rarely macron is used. Sources that mark tones with diacritics but are not accompanied by a an explanation of the representation are often ambiguous, because of the overlapping uses of macron and no mark for mid and low tones, respectively.

Numbers: Notations with numbers are common in languages with more than three levels and/or multiple contour tones. Sources vary widly as to what system of numbering they use. The two most common ones are the Chao notation or a similar system in which the highest tone corresponds to the highest number and an inverse Chao notation, in which the lowest number correspond to the highest tone. Further differences pertain to what exact levels the numbers represent. Countours are represented by joining the respective numbers, e.g. 15 for a low-high contour in a Chao notation.

The tone notations found for level tones in the sources and their standardization used in this study are summarized in Table 3.

Table 3: Tone notations in source materials

Label	Description	5	4	3	2	1
AM	all three marked	acute (á)		macron (ā)		grave (à)
MULL	mid-low grave, low underbar	acute (á)		unm. (a)	grave (à)	underbar (a̱)
MMLU	low unm., mid macron	acute (á)		macron (ā)		unm. (a)
LU	low unm., no mid	acute (á)				unm. (a)
MULM	mid unm., low macron	acute (á)		unm. (a)		macron (ā)
LM	no mid, low macron	acute (á)				macron (ā)
MULG	mid unm., low grave	acute (á)		unm. (a)		grave (à)
LG	no mid, low grave	acute (á)				grave (à)
MULB	mid unm., low underbar	acute (á)		unm. (a)		underbar (a̱)
C	Chao	5	4	3	2	1
C4	Chao with 4	4		3	2	1
C3	Chao with 3	3		2		1
IC	inverse Chao	1	2	3	4	5
IC4	inverse Chao with 4	1	2	3		4
IC3	inverse Chao with 3	1		2		3

2.2 Cuicatec

2.2.1 Concepción Pápalo Cuicatec [conc11]

The data for this variety come from Bradley (1991), who did not collect it himself but rather from other published and unpublished sources. For details about the sources refer to Bradley 1991:415. He also made reference to the dictionary by Anderson & Roque (1983) on Santa María Pápalo Cuicatec, which is said to be almost identical to Concepción Pápalo Cuicatec. Since the data sources are different and these two varieties are spoken in different towns (which are about 6h walking distance apart), I assign them to different doculects. Bradley (1991:416-418) standardized the material and uses a latin-based orthography. Tones are marked throughout.

- *x* is only described as a voiceless fricative, based on Anderson & Roque (1983) it was resolved to [ʃ]
- *d* is a voiced fricative, i.e. [ð]
- *y* is a semivowel, i.e. [j]
- laryngeal *h* (glottal stop); strictly speaking only intervocalically, in other positions it represents glottalization; for ease of of conversion, this difference has been ignored and all instances are converted to [ʔ]
- nasal vowels (marked with *n* after the vowel) are only marked in final syllables
- stress and vowel length are correlated, such that vowel length is predictable from stress; however sequences of tones are represented with two vowels, but this does not indicate length; these combinations are reduced in the conversion
- MMLU: high = acute, mid = macron, and low = unmarked
- contour tones are represented by doubling the vowel and adding the respective diacritics; I convert them to a single vowel with a following contour tone, e.g. āa = [a ˧˩]
- there is tone sandhi; surface tones are represented

- tone standardization by Bradley (1991): the two mid tones from Anderson & Roque (1983) have been merged into one mid tone because they often vary freely (cf. Section 2.2.2)

2.2.2 Santa María Pápalo Cuicatec [papa11]

The data for this variety come from an extensive dictionary by Anderson & Roque (1983). The source uses a Spanish-based orthography and the graphemes are explained only in relation to Spanish (Anderson & Roque 1983:773-793).

- *j* is most probably [x]
- *y* = [j]
- *nd*, the prenasalized stop, is distinguished from *n̲d* in compounds, which represents the sequence[n+d]
- *d* is always a fricative
- six vowels, all oral and nasal, but [ɑ] and [a] change quality due to nasalization
- *a̲* = [ɑ], *a* = [a]
- *en* (this is the nasalized *a*/[a]) = [ẽ], *e̲n* = [ẽ] (as far as I understand they are pronounced the same), but *an* (this is the nasalized *a̲*/[ɑ]) = [ã]
- IC4: there are four levels marked with numbers 4 (do), 3 (mi), 2 (sol), 1 (la)
- tone 2 and 3 vary freely in many contexts, but I left them separate because it is easier to collapse them later; the contexts are: CV(2/3)4 | CV24(2/3) | CV4C(2/3)
- based on that, tone 2/3 seem clearly distinguishable from 4, so I converted them to mid and mid-high (rather than mid-low)

2.2.3 San Juan Tepeuxila Cuicatec (1900) [tepe03]

The data for this historic variety come from Belmar (1902). The orthography used is Spanish-based but very inconsistent. Even so, the data is good enough to establish whether a cognate is present or not. IPA correspondences are based on comparison with the the modern Cuicatec varieties. Tones are marked on a few items only and they are ignored here because they are too unreliable and ambiguous.

- palatalized: *tiV* = [tʲ], *chiV* = [tʃʲ]
- *y* = [j]
- *h* = [x] (not sure)
- *ain* = [ẽ]

2.3 Triqui

2.3.1 San Andrés Chicahuaxtla Triqui [chic02 and chic12]

There are two modern sources for this variety, Good (1978) and Hernández Mendoza (2020), and a historical one from the late 19th century (Belmar 1897). The historical source is not reliable when it comes to specific phonemes, since there are many inconsistencies in writing and the author probably could not hear all of the phonemic contrasts consistently. I thus converted this as best as possible such to make it comparable to the modern sources. While

it cannot be used for reconstruction, it should still be possible to determine whether a given cognate is present in the data.

Historical Source [chic02]

- double vowels are reduced to a single vowel; vowel length is not contrastive in Triqui and it has not been written consistently anyway
- there are a few tone marks (acute, circumflex), but these are ignored in the conversion because they are too unreliable
- other equivalences where established based on comparison with the modern data, taking into account Spanish spelling-sound equivalences as well
- the data is generally not segmented into morphemes, due to all the inconsistencies

Good (1978:5, 7) uses a Spanish-based orthography. A chart of graphemes is provided, but no further explanations given.

- gottal stop is written as apostrophe, aspiration with final *j*
- *x* = [ʃ] and *y* is the semivowel [j]
- geminates are written as double consonants
- di- and trigraphs: *hu*V = [w] and *huu*V = [wː], *chr* = [ʈʂ]
- *ï* should be [ɨ] and *ë* should be [ə], comparing with Group B
- IC: tones are numbered 1 to 5, with 5 the lowest and 1 the highest
- syllables without a number correspond to tone 3, which he calls '*voz natural*' (natural voice); these will have to be added by hand for those not at the end of a word

The entries from Hernández Mendoza (2020) are already in IPA and the tones correspond the Chao notation:

- C: there are ten lexical tones, written as: 1 2 3 4 35 13 23 43 32 31
- glottalization and aspiration after vowels (with tone) cannot be currently implemented in CLTS and is thus represented by either a glottal stop or [h] as a workaround

2.3.2 San Juan Copala Triqui [copa11]

The data for this variety come from Hollenbach (1992). The orthography used is latin-based.

- retroflex consonants are indicated by *r* alone or at the end of a digraph
- *w* and *y* are glides
- three laryngeals: *h* = glottal stop, *x* = glottal spirant, i.e. [ɦ], and *"an abstract laryngeal akin to a ballistic accent. This abstract laryngeal occurs only in word-final position, and its most important phonetic manifestation is a shortening of the preceding vowel. In this sketch, a single vowel at the end of a word represents a vowel checked by this laryngeal, and a double vowel represents a word-final vowel unchecked by any laryngeal."* note: I treat this as equivalent to glottalized vowels in other Triqui varieties, otherwise there would be issues with comparison; all double vowels are converted to simple IPA ones
- C: there are eight lexical tones, namely 1 2 3 4 5 13 31 32
- most non-final syllables do not carry contrastive tone; these are left unmarked, i.e. are not followed by a number

2.3.3 San Martín Itunyoso Triqui [itun12]

The data for this variety come from Christian DiCanio's original fieldwork. The orthography used is latin-based. IPA correspondences have been established based on DiCanio (2008) and on additional explanation provided by DiCanio (p.c.).

- *y* = [j]
- *j* = [ʰ] (this appears only word-finally after vowels)
- glottal stop is written as *h*
- retroflex consonants are written as di- and trigraphs with a final *r*, e.g. *cchr* = [ʈʂː]
- geminates are written as double consonants, e.g. *yy* = [jː]
- there are pre-glottalized consonants (all simple sonorants have a glottalized counterpart), but due to CLTS limitations this is entered as glottal stop + sonorant sequences
- C: 9 lexical tones, namely 35, 4, 3, 2, 13, 43, 32, 31

2.4 Mixtec

2.4.1 Josserand 1983

The data gathered in Josserand (1983) covers 120 Mixtec varieties. They are presented in an Americanist Phonetic notation at a level between phonetics and phonology. The conventions are explained in prose in Josserand (1983:472-474). I summarize them here for convenience.

Segments

- caron = alveo-palatal fricative, e.g. *š* = [ʃ]
- *s̆* = fronted, pronounced with the blade of the tongue, which I understand as palatal, i.e [ɕ]
- *ṣ̆* = retroflex, i.e. [ʂ]
- *x* = velar fricative, i.e. [x]
- *h* = almost frictionless velar fricative (from PMx *s and not *x), i.e. [h]
- *ʉ* = [ʉ]
- sometimes an initial glottal stop is noted, but this is never contrastive and was deleted for better comparison
- nasalization on vowels is marked by a cedilla underneath the respective vowel: e.g. *ą* =[ã]
- prenasalized consononants are marked with a preposed superscript *n*, e.g. *ⁿd* = [ⁿd]
- palatalized consononants are marked with a postposed supercript *y*, e.g. *tʸ* = [tʲ]
- capital letters represent voiceless or aspirated vowels or consonants
- specific to (San Martín) Peras Mixtec [pera11]: this variety has a voiceless fricative [h], which appears only between vowels and is represented as a capital letter of the same vowel, e.g. *oOo* = [oho] (cf. also 2.4.23 for a closely related variety with the same phenomenon)

Tones Tones are not standardized and marked only for a few varieties and often only for part of the entries. They are marked with diacritics, but not according to a consistent notation and thus it is not always possible to assess which notation has been used. Table 4 provides an overview of the approximate number of entries marked for tone in Josserand (1983) and their notation (if known).[1]

[1] Additional abbreviations used in the table: A = acute, G = grave

Table 4: Overview of tone marking and notation in Josserand 1983

Mixtec Variety	Code	Josserand's Code	Entries (in %)	Notation
San Miguel Achiutla	achi11	achi	<5	A
Santa Catarina Adéquez	adeq11	adeq	0	
Alacatlatzala	alac11	alac	20-30	MULG
Alcozauca de Guerrero	alco11	alco	<5	MMLG
San Miguel Amatitlán	amat11	amat	0	
Santa María Apazco	apas11	apas	0	
Santiago Apoala	apoa11	apoa	0	
San Esteban Atatlahuca	atat11	atat	30-50	MULM
San Agustín Atenango	aten11	aten	20-30	MMLG
San Pedro Atoyac	atoy11	atoy	0	
Tepango	tepa11	ayut	20-30	MMLU
San Pedro Coxcaltepec Cántaros	coxc11	cant	0	
Chalcatongo de Hidalgo	chal11	chal	0	
San Pedro Chayuco	chay12	chap	<5	G
San Agustín Chayuco	chay11	chay	<5	G
Santiago Chazumba	chaz11	chaz	0	
San Miguel Chicahua	chic11	chic	<5	G
San Juan Coatzospan	coat11	coat	0	
Coicoyán de las Flores	coic11	coi	30-50	MMLG
San Juan Colorado	colo11	colo	0	
Cosoltepec	coso11	cos	<5	A
San Cristóbal Jamiltepec	cris11	cris	0	
Santa Cruz de Bravo	brav11	cruz	<5	G
Santa Ana Cuauhtémoc	cuau11	cuau	0	
Cuyamecalco Villa Zaragoza	cuya11	cuya	0	
San Juan Diuxi	diux11	diux	0	
San Martín Duraznos	dura11	durz	20-30	MMLG
Santa Catarina Estetla	este11	este	0	
San Francisco de las Flores	flor11	flor	0	
San Antonio Huitepec	huit11	huit	0	
Santa Cruz Itundujia	itun11	itun	0	
San Miguel Ixtapam	ixta13	ixtp	<5	G
Santiago Ixtayutla	ixta11	ixty	0	
San Francisco Jaltepetongo	jalt11	jalt	0	
Santiago Jamiltepec	jami11	jam	0	
San Jerónimo Xayacatlán	xaya11	jer	0	
Santa María Jicaltepec	jica11	jict	0	
San Pedro Jicayán	jica12	jicy	0	
San Pedro Jocotipac	joco11	joco	0	
Santiago Juxtlahuaca	juxt11	juxt	<5	G
San Agustín Monte Lobos	mont14	lobo	0	
San Lorenzo	lore11	lor	0	
Metlatónoc	metl11	metl	90-100	MMLG
El Rosario Micaltepec	mica11	mic	0	

Mixtec Variety	Code	Josserand's Code	Entries (in %)	Notation
San Miguel El Grande	gran11	mig	30-50	MULM
San Sebastián del Monte	mont12	mont	0	
Santa Cruz Nundaco	nund11	ndac	0	
Santiago Nundiche	nund12	ndi	0	
Santo Domingo Nundó	nund13	ndo	0	
San Antonio Nduayaco	ndua11	ndua	0	
San Jorge Nuchita	nuch11	nuch	20-30	MMLG
Santo Domingo Nuxaa	nuxa11	nuxa	0	
San Andres Nuxiño	nuxi11	nuxi	0	
Santiago Nuyoo	nuyo11	nuyo	90-100	MMLG
San Juan Ñumí	numi11	ñumi	0	
Santo Tomás Ocotepec	ocot11	oco	90-100	MMLG
Proto-Mixtec	prom01	p-mixtec	0	
San Mateo Peñasco	pena12	peña	0	
Santa María Peñoles	peno11	peño	75	LG
San Martín Peras	pera11	pera	20-30	MMLG
San Miguel Piedras	pied11	pied	<5	G
Pinotepa de Don Luís	pino12	pinL	<5	A
Santiago Pinotepa Nacional	pino11	pinN	<5	A
San Miguel Progreso Mixtec	prog11	prog	0	
San Francisco Sayultepec	sayu11	sayu	<5	A
San Martín del Estado	esta11	silM	0	
San Jerónimo Progreso	prog12	silP	<5	G
San Jose Sinicahua	sini11	sinc	0	
San Mateo Sindihui	sind11	sind	0	
San Jeronimo Sosola	soso11	soso	0	
San Juan Tamazola	tama11	tamJ	0	
Santa María Tataltepec	tata11	tata	0	
San Sebastian Tecomaxtlahuaca	teco11	teco	30-50	MMLG
San Juan Teita	teit11	teit	0	
San Pedro y San Pablo Teposcolula	tepo11	tepo	0	
San Pedro Tidaa	tida11	tida	0	
Santiago Tilantongo	tila11	til	0	
Santa Catarina Tlaltempan	tlal11	tlal	<5	A
Santa María Totoltepec de Guerrero	toto11	tot	0	
San Pedro Tututepec	tutu12	tut	<5	G
Santa Lucía Monteverde	mont13	verd	0	
Xayacatlán de Bravo	xaya11	xay	90-100	MMLG
San Pedro Yucunama	yucu14	ynam	0	
Santiago Yosondua	yoso	yoso	0	
San Bartolomé Yucuañe	yucu11	yuca	0	
Santa María Yucuhiti	yucu12	yuci	90-100	MMLG
Santa María Yucuñuti de Benito Juárez	yucu13	yucñ	20-30	MMLG
Santa María Yucunicoco	yucu15	yuco	90-100	MMLG
Santa María Yucuquimi de Ocampo	yucu16	yucq	0	
San Juan Yuta	yuta12	yuta	0	
Yutanduchi de Guerrero	yuta	yutn	0	

Mixtec Variety	Code	Josserand's Code	Entries (in %)	Notation
Santa María Zacatepec	zaca11	zac	0	
Zapotitlán Palmas	zapo	zap	<5	G

2.4.2 Dürr 1987

The data presented in Dürr (1987) covers 17 Mixtec varieties. All data have been aggregated from published sources (cf. Dürr 1987:36-37 for the complete list) and standardized in Americanist Phonetic notation. The segmental notation is thus essentially the same as in Josserand (1983) and the reader is referred to Section 2.4.1 for more details. A few subphonemic details such as nasalization have been included for better comparison. The cognate sets are arranged in the same way as in Josserand (1983). The tones are standardized and marked as follows (Dürr 1987:p.57 footnote 16):

- AM: high = acute accent, mid = macron, low = grave accent
- contour tones: high-low = circumflex, mid-low = macron-grave, low-high = caron, mid-high = macron-acute

A few varieties have some special notation:

- Diuxi: tones are given according to Daly (1978); the apostrophe indicates *"non-predictable word-final stress, which is a feature of tone"*
- Peñoles: circumflex = modified low tone, caron = modified low tone
- tone notation in Cahuatache is not completely clear from the source material

2.4.3 Abasolo del Valle Mixtec [abas11]

This variety is spoken in a diaspora comunity in the state of Veracruz. The data come from Galindo Sánchez (2009). The author claims that the variety is identical to that spoken in San Juan Mixtepec. The orthography is latin-based without further explanations given (Galindo Sánchez 2009:12, 18-23). Unclear IPA correspondences are taken over from Mixtepec Mixtec.

- *x* = [ʃ] and y = [ʒ]
- *ku*V = [k^{w}]
- MULG: high = acute, mid = unmarked, low = grave (three tonemes)

2.4.4 Alacatlatzala Mixtec [alac11]

The data for this variety comes from five sources: Josserand (1983), Dürr (1987), Zylstra (1991), Anderson (2006) and Zylstra (2012). Tones are marked in all of these sources, but not for all entries in Josserand (1983). As detailed below, the orthographies used across the five sources are different and produce certain conflicts and overlaps (especially in tone notation). For correct conversion, the doculect was initially split into three sets: Set A = Anderson (2006) and Zylstra (2012), Set B = Zylstra (1991), and Set C = Josserand (1983) and Dürr (1987) (see Sections 2.4.1 and 2.4.2 for more details).

Set A Anderson (2006:viii) and Zylstra (2012:13) both use a Latin-based system

- digraphs: ti*V* = [t^{j}], ku*V* = [k^{w}]
- *y* is a semivowel, i.e. [j]
- *x* = [ʃ]

- MULB: high = acute, mid = unmarked, low = underbar

Set B Zylstra (1991:8) is also a Latin-based system but with different grapheme choices and tone notation

- but *x* = [x] (*sh* = [ʃ])
- glottal stop as *h*
- MMLU: high = acute, mid = macron, low = unmarked

2.4.5 Alcozáuca de Guerrero Mixtec [alco11]

The data for this variety come from two sources: Josserand (1983) and Swanton & Mendoza Ruíz (forthcoming). Tone is marked for a few entries only in Josserand (1983) with diacritics.

- *c* = [c] (this consonant is unique to Alcozauca Mixtec and transcribed as [t] in Josserand (1983))
- *b* = [β]; as in other varieties, this is realized as [b] word-initially, but [β] word-medially, but since this kind of allophony is not represented in other sources and for better comparison, I chose [β] as the only representation
- C4: four level tones and three contours, namely 1 = low, 2 = mid-low, 3 = mid, 4 = high, 14 = low-high, 41 = high-low, 141 = high-low-high

2.4.6 San Esteban Atatlahuca Mixtec [atat11]

The data for this variety comes from three sources Alexander (1980), Josserand (1983), and Dürr (1987). Tones are marked in all sources, but not on every entry in Josserand (1983).

Alexander (1980) uses a spanish-based system

- *jn* = [n̪]
- *x* = [ʃ], *y* = [ʒ] (not specified, but most likely comparing with other sources)
- *j* is not further explained apart from being pronounced like Spanish, but comparing with Group B and C most likely it is [h]
- glottal stop is written as *h*
- MULM: high = acute, mid = unmarked, low = macron

2.4.7 La Batea Mixtec [bate11]

The data for this variety was shared with us by collaborator and native speaker Yésica Ramírez (MICOP). The data were collected between 2016-2020 in Oxnard and Santa Barbara, California as part of a collaboration between the Linguistics department of the University of California Santa Barbara and the Mixteco/Indígena Community Organizing Project in Oxnard, California (Hernández Martínez et al. 2021). The orthography used is Latin-based.

- there are palatalized consonants: *ti*V = [tʲ], *tsi*V = [tsʲ]
- *y* = [j]
- *x* = [ʃ]
- MULG: there are five tonemes: high = acute, mid = unmarked, low = grave, high-low = circumflex, low-high = caron

2.4.8 Cahuatache Mixtec [cahu11]

The data for this variety originally comes from Schultze-Jena (1938), a collection of early travel accounts. It is reproduced in Josserand (1983) and Dürr (1987). Although a valuable resource, the interpretation of the material is not easy, especially regarding the tone notation. Therefore, I only use the data as reproduced in the later collections, not from the original. Tones are only marked in Dürr (1987).

- *y* is resolved to [j], based on the information on other varieties spoken in Guerrero

2.4.9 Chalcatongo de Hidalgo Mixtec [chal11]

The data for this variety comes from three sources: Josserand (1983), Macaulay (1996), and Swanton & Mendoza Ruíz (forthcoming). Tone does not seem to be marked in the Josserand (1983) entries.

Macaulay (1996:19) uses an Americanist notation, with a few idiosyncrasies:

- *b* is described as a voiced stop and written as such, although it is pronounced either [^{m}b] or [β] depending on the position
- nasalization is marked with a tilde
- MULG: high = acute, mid = unmarked, low = grave

Swanton & Mendoza Ruíz (forthcoming) write in IPA:

- *y* is a vowel [ʉ] in one entry (LAZY)
- C3: low = 1, mid = 2, high = 3

2.4.10 San Agustín Chayuco Mixtec [chay11]

The data for this variety comes from three sources Pensinger et al. (1974), Josserand (1983), and Dürr (1987). Pensinger et al. (1974:137) uses a Spanish-based orthography and tones are marked for a few entries only. The same is true for Josserand (1983), while those from Dürr (1987) are marked for tone throughout. The Josserand (1983) entries could be from a different, nearby village, given that there are slight but regular segmental differences, such as [tʃ] where the others have [s].

- *h* = [ʔ]
- *z* in one word only 'child' is 'pronounced like in Spanish'; based on the other sources, this must mean pronounced like Castillian Spanish, i.e. as [θ]; in other words it seems to be [s]
- *x* = [ʃ] (only in Pensinger et al. 1974)

2.4.11 San Juan Coatzóspam Mixtec [coat11]

The data for this variety comes from three sources: Josserand (1983), Dürr (1987), and Small (1990). Tones are not marked in Josserand (1983), but the other two sources.

Small (1990:268) uses a latin-based system:

- LM: there are three tones, namely high = acute, low = macron, and a rare high-low glide written with circumflex
- *"Unaccented syllables do not carry contrastive tone; they assimilate to the tone of the following syllable. Tone is not marked on such syllables in this sketch"*, i.e. unmarked really means toneless here

- this variety has extensive tone sandhi, surface tones are written
- *d* is the voiced fricative [ð], there's also a palatalized version *dy* = [ðʲ]
- *h* is the glottal stop

2.4.12 San Juan Colorado Mixtec [colo11]

The data for this variety comes from Josserand (1983) and Stark et al. (1986). Tones are only marked in the latter source.

Stark et al. (1986:145) uses a Spanish-based orthography:

- *x* = [ʃ], cf. *"x representa un sonido semejante al de la ch pero pronunciada sin pegar la lengua a los alveolos"*
- *j* = [h] and y = [j]; it is only mentioned that they are pronounced like Spanish but comparing to Group B these values make the most sense
- MULG: high = acute, mid = unmarked, low = grave

2.4.13 San Juan Diuxi Mixtec [diux11]

The data for this variety comes from three sources: Josserand (1983), Dürr (1987), and Kuiper & Oram (1991). Tones are marked in the latter two sources. They differ quite drastically, however, because in Dürr (1987) they are presented according to the reanalysis of Daly 1978, in Kuiper & Oram (1991) this is true for some sections but not others. A detailed report on the differences and how they could be reconciled are outside the scope of the present study.

Kuiper & Oram (1991:186) use a latin-based system

- *y* is a voiced fricative [ʒ]
- *x* = [x], *sh* = [ʃ]
- laryngeal h (glottal stop)
- two different systems are used, but fortunately the markings do not clash; the tones are not directly comparable though
- LU: sections by Oram rely on the 'Pike-Oram' system in which underlying tones are written as high = acute, low = unmarked
- MMLU: sections by Kuiper note surface tones, namely high = acute, mid = macron, low = unmarked, high-low = circumflex

2.4.14 San Martín Duraznos Mixtec [dura11]

The data for this variety comes from two sources: Josserand (1983) and the ongoing documentation project of Sandra Auderset and native speaker Carmen Hernández Martínez. Tones are only marked on a few entries in Josserand (1983), but throughout in the newer source. In the current documentation project, we use a latin-based orthography in which the mid tone is unmarked. This variety exhibits a rare split between alveo-palatal and palato-alveolar sibilants and affricates, which is not reflected in Josserand (1983).

- alveo-palatals: *sh* = [ʃ], *ch* = [tʃ], and *nch* = [ⁿdʒ]
- palato-alveolars: *x* = [ɕ], *tx* = [tɕ], and *ntx* = [ⁿdʑ]
- *y* = [ʒ]
- *ku*V = [kʷ]
- glottal stop is written as apostrophe

- MULG: there are three tones, namely high = acute, mid = unmarked, low = grave

2.4.15 San Marcos de la Flor Mixtec [flor12]

The data for this variety was shared with us by collaborator and native speaker Moisés (MICOP). The data were collected between 2019-2020 in Oxnard and Santa Barbara, California as part of a collaboration between the Linguistics department of the University of California Santa Barbara and the Mixteco/Indígena Community Organizing Project in Oxnard, California (Hernández Martínez et al. 2021). This variety is closely related to Piedra Azul Mixtec and has the same phoneme inventory. It also uses the same writing system, so refer to Section 2.4.23 for details.

2.4.16 San Miguel El Grande Mixtec [gran11]

The data for this variety comes from two sources: Josserand (1983) and Dürr (1987). Tones are marked in both, but not for all entries in Josserand (1983).

2.4.17 Santo Domingo Huendío Mixtec [huen11]

The data for this variety come from Becerra Roldán (2015). The author uses a latin-based system and tones are marked throughout with diacritics.

- *x* = [ʃ] and y = [ʒ]
- *j* = [x]
- *ku*V = [k^w]
- MULG: high = acute, mid = unmarked, low = grave

2.4.18 Santiago Jamiltepec Mixtec [jami11]

The data for this variety come from two sources: Josserand (1983) and Johnson (1988). Tones are only marked in the latter.

Johnson (1988:18) uses a latin-based system:

- *y* is the semivowel [j]
- *x* = [x]
- there are palatalized consonants: ty = [t^j] and ndy = [$^nd^j$]
- MMLU: high = acute, mid = macron, and low = unmarked

2.4.19 El Jicaral Mixtec [jica11]

The data for this variety was provided by JN Martin. The orthography used is Latin-based and tones are marked throughout.

- *x* = [ʃ]
- *y* = [j]
- *ku*V = [k^w]
- there are palatalized consonants: *ti*V [t^j] and *nti*V [$^nd^j$]
- MULB: there are five tonemes: high = acute accent, mid = unmarked, low = underbar, low-high = caron, high-low = circumflex

2.4.20 San Sebastián del Monte Mixtec [mont12]

The data for this variety come from two sources: collaborator and native speaker Juvenal Solano (MICOP, Oxnard CA) and Josserand (1983). Tones are not marked in Josserand (1983). The data from collaborator Juvenal Solano were collected between 2016-2020 in Oxnard and Santa Barbara, California as part of a collaboration between the Linguistics department of the University of California Santa Barbara and the Mixteco/Indígena Community Organizing Project in Oxnard, California (Hernández Martínez et al. 2021). In this source tones are marked throughout.

Solano 2020:

- *x* = [ʃ] and *y* = [j], *j* = [h]
- there are palatalized consonants: *ty* = [tʲ] and *ndy* = [ⁿdʲ]
- MULG: three tones: high = acute, mid = unmarked, low = grave

2.4.21 Santo Tomás Ocotepec Mixtec [ocot11]

The data for this variety comes from three sources: Josserand (1983), Dürr (1987) and Alexander (1988). Tones are marked in all three sources, although there are a few entries in Josserand (1983) for which tone is not marked. Otherwise, Josserand (1983) uses the same system as Dürr (1987).

Alexander (1988:170) uses a latin-based system:

- *y* is a voiced fricative, i.e. [ʒ], *x* = [x]
- glottal stop is written as *h*
- three tones: high = acute, mid = macron, and low = unmarked

2.4.22 Magdalena Peñasco Mixtec [pena11]

The data for this variety comes from an extensive, recently published dictionary by Hollenbach (2017) using a latin-based system. The graphemes are listed in Hollenbach (2017:xix-xxi). Information for the corresponding IPA sounds has also been taken from the grammar by the same author (Hollenbach 2013:9-14, 16-18).

- *d* = [ð]
- *j* = [h]
- *x* = [ʂ]
- *y* = [ʒ] ([ʃ] at the beginning of words)
- digraphs: *ku*V = [kʷ], *nd* = [ⁿd] and *tn* = [tⁿ]
- there are palatalized consonants: *ti*V = [tʲ] and *si*V = [sʲ] (the latter does not appear in the collected entries)
- there are three tones: high, mid, low
- in the dictionary, these are not marked on the entry but given in square brackets after the entry, represented with letters; for better comparison, I have changed this so that the entries are now directly marked for tone with high = acute, mid = unmarked, low = grave (MULG)

2.4.23 Piedra Azul Mixtec [pied12]

The data for this variety was provided by linguist and native speaker Gabriel Mendoza with Simon L. Peters (Mendoza & Peters 2020). Our collaboration is part of the MILPA initiative, a collaboration between the Linguistics department of the University of California Santa Barbara and the Mixteco/Indígena Community Organizing Project in Oxnard, California (Hernández Martínez et al. 2021). The orthography used is latin-based and tones are marked throughout.

- *j* = [h]
- *x* = [ʃ]
- *y* = [j]
- *ku*V = [k^w]
- palatalized consonants: *ki*V = [k^j], *ti*V = [t^j], *tsi*V = [ts^j], *ntsi*V = [$^n ts^j$]
- MULG: five tonemes: high = acute, mid = unmarked, low = grave, high-low = circumflex, low-high = caron

2.4.24 San Jerónimo Progreso Mixtec [prog12]

The data for this variety come from three sources Josserand (1983), Dürr (1987), and Shields (1988). Tones are marked in the latter two sources. The IPA correspondences for Shields (1988) were identified based on the phonology outline provided in North & Shields (1977). Note that the sources refer to this variety as Silacayoapan Mixtec, using the name of the municipality. They all specify, though, that the data was gathered in the town of San Jerónimo Progreso.

Shields (1988:312) uses a latin-based orthography:

- *h* = [ʔ]
- *x* = [h]
- *y* is a voiced fricative, i.e. [ʒ]
- palatalized consonants: ky = [k^j], xy = [h^j], kwy = [k^{wj}]
- MMLU: high = acute, mid = macron, and low = unmarked (surface form tones)

In Josserand (1983) and Dürr (1987), some graphemes are used differently:

- *h* represents [h] and *x* = [x]
- unmarked vowels are those that are not marked for tone

2.4.25 Tepango Mixtec [tepa11]

The data for this variety come from three sources Josserand (1983) (see 2.4.1), Dürr (1987) (see 2.4.2), and Hills (1990). All three sources refer to the variety as 'Ayutla Mixtec', referencing the municipality that Tepango is part of. They all specify that the data was gathered with speakers in Tepango. This is one of only a few Mixtec varieties that retains final glottal stop.

Hills (1990:8) uses a latin-based system:

- palatalized consonsants: *ty* = [t^j], *ny* = [n^j]
- glottal stop is *h*
- *y* is semivowel [j]
- MMLU: high = acute, mid = macron, and low = unmarked

In Josserand (1983) and Dürr (1987), some graphemes are used differently:

- *y* is semivowel [j], but it appears twice at the end of word in Josserand (1983) and in both cases Hills (1990) has [ũ]; because of this and because Mixtec words cannot end in a consonant, I convert it to [i] in this position
- *ɛ* is neutralized to [e]

2.4.26 San Pedro y San Pablo Teposcolula Mixtec [tepo01]

For this variety there are historical records in the form of a colonial era vocabulary by de Alvarado (1962 [1593]). This source uses a Spanish-based writing system and tones are not marked. The spelling is not very consistent, but a good baseline of equivalences can be established based on comparisons with modern varieties. In addition, the variety is also documented in Josserand (1983), who standardized and interpreted the entries to some degree, which further clarifies some of the orthographical choices.

- *dz* = [ð] (this is very consistent)
- *d* = [^{n}d] at the beginning of words, in the middle the prenasalization is written as *nd*
- the glottal stop is represented by *h*, but this grapheme is also used in combinations for other sounds, such as *hu* = [w] and occasionally in the middle of a word = [ʔ w], inital *h* seems to be purely graphical (like in Spanish)
- nasalization of vowels is not marked

2.4.27 Tlahuapa Mixtec [tlah11]

The data for this variety come from collaborator and native speaker Griselda Reyes Basurto (MICOP, Oxnard CA). The data were collected between 2016-2020 in Oxnard and Santa Barbara, California as part of a collaboration between the Linguistics department of the University of California Santa Barbara and the Mixteco/Indígena Community Organizing Project in Oxnard, California (Hernández Martínez et al. 2021). Reyes Basurto (2020) uses a latin-based systems and tones are marked throughout.

- palatlized consonants: *ty* = [t^{j}]
- *x* = [ʃ] and *y* = [j]
- MULG: high = acute, mid = unmarked, low = grave, high-low = circumflex, low-high = caron

2.4.28 Xochapa Mixtec [xoch11]

The data for this variety come from Stark et al. (2013), who use a latin-based orthography. Spanish loans, however, are written in Spanish orthography Stark et al. (2013:85-90).

- prenasalized: *nd* = [^{n}d]
- *ku*V = [k^{w}]
- *x* = [ʃ] and *y* = [j]
- there are four level tones: high = acute, mid = unmarked, mid-low = grave, low = underbar, high-low = circumflex, low-high = caron

2.4.29 Yoloxochitl Mixtec [yolo12]

The data for this variety come from Josserand (1983) and Amith & Castillo García (n.d.). In Josserand (1983) only two items are marked for tone with acute and grave. In the latter source tones are marked throughout with numbers. The orthographies are different, but do not clash. Amith uses a latin-based system:

- prenasalized: *nd* = [ⁿd]
- *x* = [ʃ] and *y* = [j]
- palatalized consonants: *ti*V = [tʲ]
- here *w* = [β] in Josserand (1983)
- C4: there are four level tones: low = 1, low-mid = 2, mid = 3, high = 4

2.4.30 Santiago Yosondúa Mixtec [yoso11]

The data for this variety comes from two sources: Josserand (1983) and Farris (1992).

Farris (1992:8) uses a latin-based system:

- *x* = [x] and *y* = [j]
- palatalized consonants: *shy* = [ʃʲ]
- preaspirated nasals: *Nn* = [ʰn], *Ññ* = [ʰɲ]
- MMLU: high = acute, mid = macron, and low = unmarked
- there is extensive tone sandhi, but the tones are written as underlying (i.e. before the application of any sandhi rules)

2.4.31 San Pedro Yosoñama Mixtec [yoso14]

The data for this variety come from Gittlen (2016). The grammar refers to the variety as 'Mixteco del norte de Tlaxiaco' (Mixtec of northern Tlaxiaco), but details that the data were collected in the town of San Pedro Yosonama. A latin-based system is used, but tones are unfortunately not marked. The orthography is explained only in relation to Spanish (Gittlen 2016:3-10).

- *x* is [ʃ]
- *y* is [ʒ]
- glottal stop is marked by apostrophe

2.4.32 San Andrés Yutatio Mixtec [yuta13]

The data for this variety come from an extensive recent dictionary Williams et al. (2017). The orthography used in this source is latin-based and tones are marked throughout. It is presented in a chart with few further explanations (Williams et al. 2017:x-xiv). However, a more detailed description is provided in the pedagogical grammar by the same author (Ferguson de Williams 2007:9-20) and I base my analysis on this grammar.

- *d* = [ð]
- *j* = [h]
- *y* = [j] 'se pronuncia como suena la ll de calle al estilo oaxaqueño' (it is pronounced like the ll of street in the Oxacan style)
- *x* = [ʃ]

- glottal stop is apostrophe
- MULB: high = acute, mid = unmarked, low = underbar

2.4.33 Yucuquimi de Ocampo Mixtec [yucu16]

The data for this variety comes from two sources: Josserand (1983) and Swanton & Mendoza Ruíz (forthcoming). Tones are marked only in the latter with numbers.

- *b* = [β]
- *y* is used as a vowel, most probably [ʉ]
- *kɥ* in Swanton & Mendoza Ruíz (forthcoming) corresponds to [kʷ]
- C3: low = 1, mid = 2, high = 3

2.4.34 Yucunani Mixtec [yucu17]

The data for this variety was shared by collaborator and native speaker Jeremías Salazar. The orthography used is latin-based **?** and tones are marked throughout.

- prenasalized: *nt* = [ⁿd], *nch* = [ⁿtʃ]
- *ku*V = [kʷ]
- *x* = [ʃ] and *y* = [j]
- MULG: high = acute, mid = unmarked, low = grave, high-low = circumflex, low-high = caron

2.4.35 Santa María Zacatepec Mixtec [zaca11]

The data for this variety comes from three sources: Josserand (1983), Swanton & Mendoza Ruíz (forthcoming), and Towne (2011). Tones are marked in Swanton & Mendoza Ruíz (forthcoming) with numbers and on a few entries in Towne (2011) with diacritics. The orthographies used are different but do not clash. This variety is very important for reconstruction because it preserves final glottal stop. However, this feature is missing from a lot of entries in Towne (2011).

- *y* = [j]
- *ku*V =[kʷ]
- *x* = [ʃ] (x does not appear in the other two sources, so there is no overlap)
- there is one entry with *j* = [h], which has to be adjusted by hand (because it's used as [j] in the other sources)
- Towne (2011) uses MMLB: high = acute, mid = macron, low = underbar; but only marked for (segmental) homophones
- Swanton & Mendoza Ruíz (forthcoming) uses C3: low = 1, mid = 2, high = 3

3 Cognate coding

In this section we summarize the main principles we followed for assigning the cognates to sets and discuss the more difficult cases and our approach to them.

There is only one study dealing exclusively with the reconstruction of proto-Mixtecan (PMx), which is now over 60 years old, namely that by Longacre (1957). His work was based on four Mixtec varieties, one Cuicatec variety, and three Triqui varieties. While a very valuable first step in the reconstruction of PMx, many of the proposed correspondences and sound laws need to be revised now that we know so much more about these languages. We consulted Longacre (1957) as a reference point, but did not directly base our cognate judgments on his work for the reasons mentioned above. Mixtecan was also mentioned in earlier work on larger groupings, e.g. in the discussion between Swadesh (1960) and Longacre (1961) on the status of Mixtecan within Otomanguean.

The Mixtec branch, comprising the largest group of languages, has received most attention with respect to reconstruction. Several scholars have provided and refined reconstructions and sound changes over the years, namely Mak & Longacre (1960), Bradley & Josserand (1982), Josserand (1983), Dürr (1987), and most recently Swanton & Mendoza Ruíz (forthcoming). Josserand (1983) provides 188 proto-forms and provides sound changes and correspondences for vowels across 120 varieties. Her work remains the most extensive to date. Dürr (1987) focused on tone correspondences and tone reconstruction. He provides 110 proto-forms with ton and data from 17 varieties. Both the proto-forms and varieties overlap to a great extent with those of Josserand (1983). More recently Swanton & Mendoza Ruíz (forthcoming) provide an overview of the diachrony tone in Alcozauca Mixtec. However, they also provide data from three other Mixtec varieties, as well as 84 proto-forms with tone. We drew mainly on these three studies for cognate assignments within Mixtec.

Proto-Triqui has been reconstructed by Matsukawa (2005), but there are several issues with this work. For one, it ignores the geminate consonants in Itunyoso Triqui which contain important clues about initial syllables that allows linking them with proto-Mixtec forms (add citation). Furthermore, the Chicahuaxtla Triqui data used for reconstruction is based on field-work conducted during a very short period of time and so comes with some inconsistencies. We have thus opted not to incorporate this study or the data presented there.

Unfortunately, Cuicatec is not only underdocumented but also understudied with respect to its diachrony. Consequently, no reconstruction of proto-Cuicatec is available, nor has there been any work after Longacre (1957) refining the sound laws between Cuicatec and the two other branches.

To sum up, there is no comprehensive, agreed upon reference work for Mixtecan reconstructions and sound laws. This means that the coding of the cognate sets is necessarily tied to our current understanding of the sound correspondences within and between the three branches. We have opted for a conservative approach. This means that in cases where we were unsure if two or more entries are cognate or not, we assigned them to different cognate sets.

3.1 Tonal derivation

Since all Mixtecan languages make extensive use of lexical and grammatical tone, the question arises of how to treat forms that are derived from others by tone. An example from San Juan Coatzospam Mixtec is 'black' $tu^{5}\tilde{u}^{1}$, which is derived from 'charcoal' $t\tilde{u}^{1}\tilde{u}^{5}$. Tonal derivation of adjectives from nouns is common among Mixtec and Triqui varieties (Hinton et al. 1991).

With respect to cognate assignments, we treat tonal derivation the same as segmental derivation. That is, we assign the tonal derivational morpheme its own cognate ID and represent it in the appropriate spot, as a prefix. The only difference to segmental derivation is that the toneme cannot be visually segmented in the same an affix (sometimes) can be. An example of segmental and tonal derivation is presented below.

Segmental derivation in the annotation of cognate sets (Magdalena Peñasco Mixtec):

concept	entry	cognates
'tail'	*su³ʔma¹*	635
'scorpion'	*ti¹+su³ʔma¹*	9+635

Tonal derivation in the annotation of cognate sets (San Juan Coatzospam Mixtec):

concept	entry	cognates
'charcoal'	*tũ¹ũ⁵*	93
'black'	*tu⁵ũ¹*	1044+93

3.2 Broad vs. fine cognate assignments

Given the issues outlined above (Section 3), we annotated the cognate sets in two different ways: a broad analysis and a more fine-grained one. The main difference between the two analyses is that the fine-grained one takes into account potentially 'irregular' or 'unexpected' reflexes, while the broad ones ignores those. The idea behind this is that these variant reflexes could carry philogenetic signal, that is they could reflect shared innovations rather than parallel ones and thus are potentially useful for subgrouping. This is an open question, though, that we want to be able to address in an empirical way rather than prejudging the matter. This is why we opted for annotating the cognate sets in two ways, which allows to run the models on both to assess the potential influence on the outcomes.

3.2.1 L-alternations in Mixtec

In Mixtec, we sometimes find an alternation between [l] and dental fricatives or sibilants. That is, in some cognate sets, a group of Mixtec varieties will exhibit an [l] where others show [s/ð] or more rarely [ⁿd]. Note that the lateral is rare in all of Mixtecan. This is not a regular sound change in the strict sense, because it varies from variety to variety which cognate sets are affected – if any at all. Previous studies have usually treated this [l] as a Mixtec innovation and reconstructed it to proto-Mixtec (Longacre 1957, Bradley & Josserand 1982, Josserand 1983). We agree with the assessment of [l] as an innovation, but not one at the proto-Mixtec level but in individual varieties (and perhaps subgroups). Reconstructing the dental fricative/sibilant allows us to connect the proto-Mixtec forms to Cuicatec and Triqui reflexes in a more straightforward way. Furthermore, treating [l] as a language- or subgroup-specific innovation rather than a proto-Mixtec one explains its patchy distribution better. We simply have to assume that it did not diffuse throughout the whole lexicon. Regarding the origin of this innovation, it is possible that there were two phenomena at work: an aerodynamically motivated sound change (Ohala 1983, 1989), and a word formation process involving phoneastemes. The sound change giving rise to the l-form(s) in the first place could be envisaged as follows: proto-Mixtecan */ð/ > Proto-Mixtec [θ s/_a (loss of voicing) > [l]/_a (maintenance of voicing but loss of constriction) (add citation by DiCanio). Another proposed correspondence that could have

Table 5: Cognate sets with l-forms with counts and reflexes in other languages[2]

Concept	Affected Set(s)	Mixtec [l]	in %	Mixtec other	Cuicatec	Triqui
BIRD	49	38	29	[s/ð]	[ð]	[t]
FROG	263	33	35	[s/ð]	NA	DS
URINE	826	12	26	[s/ð]	DS	[r]
TAIL/SCORPION	635	7	4	[s/ð]	[ⁿd]	DS
EAR	194	4	3	[s/ð]	DS	DS
PUS	574	19	16	[ⁿd]	[j]	[kː]
SMOOTH	660	7	9	[ⁿd]	DS	[n]
BEAN	24	1	1	[ⁿd]	[n]	[r]
TOMATO	775	1	1	[n]	DS	[r]
SQUIRREL	698	1	2	[ɲ]	DS	DS
RABBIT	578	12	57	ø	DS	DS
SMALL	669	4	22	[tʃ]	DS	DS
CAT	87, 88	9	13	[t]/[st]/[tʃ]	[tʃ]	[l]

had an influence on the l-alternation is Marlett's (1992) proposal that [ⁿd] is just an allophone of /n/ before oral vowels. Since in some sets we observe an alternation between [ⁿd] and [l] and there are apparent cases of proto-Mixtec *l turning into [n] before nasal vowels...

A summary of cognate sets that exhibit this alternation is provided in Table 5 and geographical overview in Figure 1. There is no clear spatial pattern, but that does not exclude that these forms are relevant for subgrouping. We therefore chose to annotate them separately in the fine-grained cognate assessments (cf. 3.2. From Table 5, we see that the l-alternation appears in quite a few animal terms, as well as body parts, bodily secretions, and vegetables. Further research will have to show whether those are the only semantic domains affected and to what extent semantics play a role in the diffusion of the l-forms. In terms of the alternation itself, the top five sets show the expected l-s/ð-alternation, but there are also three forms in which l alternates with [ⁿd] and a further five sets with varying dental/alveolar consonants. At this point, we cannot say whether these are all part of the same innovation, or whether separate processes should be assumed here. We do want to point out that CAT should probably be considered separately, since it is known to be a Mesoamerican *Wanderwort* (Kiddle 1964).

3.2.2 Fossilized verb morphology

Some forms exhibit fossilized verb morphology, i.e. consonant alternations that look irregular because they come from lexicalizations of preverbal formatives. In Mixtec, this is most often involves an initial [k] alternating with another initial consonant. This [k] marks potential mood and is still present on many verbs, although no longer productive. Also common are fossilizations of the iterative marker [ⁿd] as the initial consonants. We find this in set 388 for LIGHTNING, where the initial consonant varies between [t], [ⁿd] and [k]. The majority of Mixtec languages have exhibit a reflex of PM *ta^5xa^5, but Tlahuapa Mixtec also has a form $ka^3ʃã^3$ and San Sebastian del Monte Mixtec has a form $^nda^3sa^3$. We interpret those as lexicalized irrealis and iterative forms, respectively. In the broad annotation, we do not distinguish them from the other forms for 'lightning', but in the fine-grained annotation we assign them to different sets. Triqui and Cuicatec varieties do not show reflexes of this cognate set.

[2]Abbreviations used in the table: DS = different cognate set(s), NA = not attested

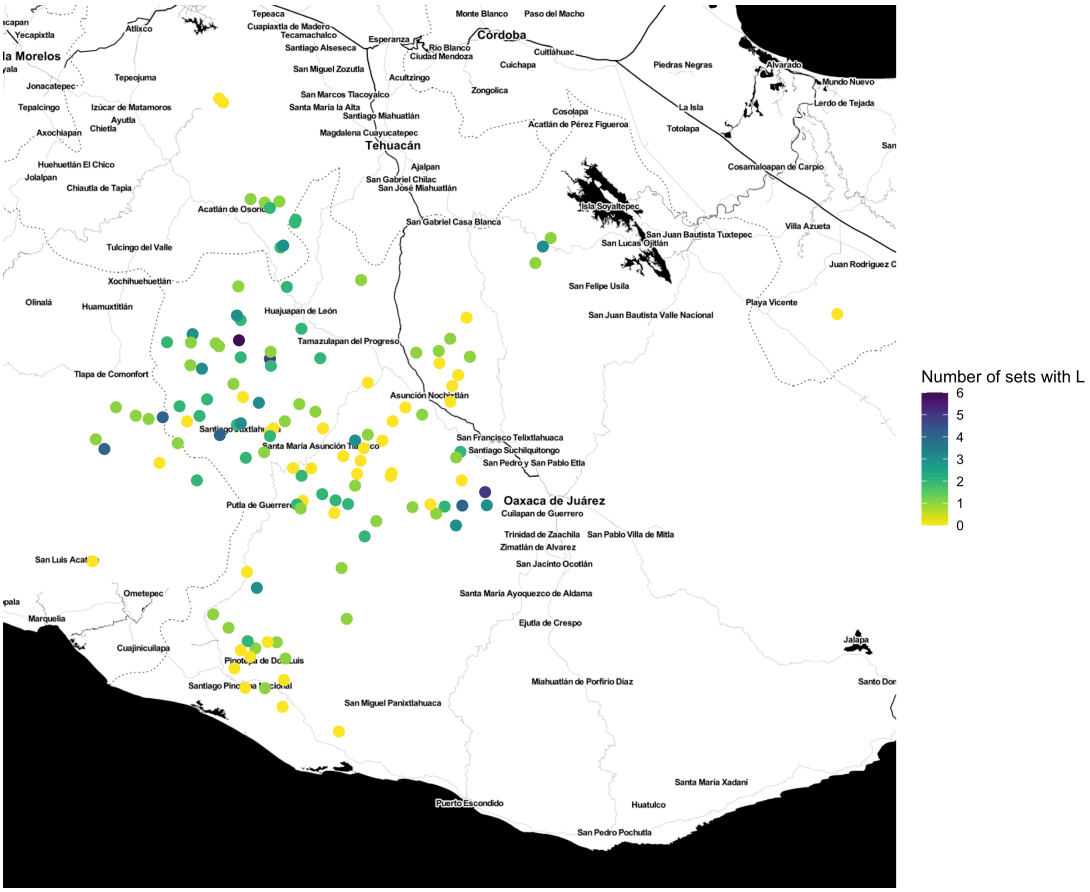

Figure 1: Number of cognate sets with L-forms in Mixtec languages

3.2.3 Other alternations resulting from fossilized prefixes

Some forms that are derived from others show consonant alternations that seem irregular, but are not fossilized verb morphology (cf. 3.2.2). These can probably be explained by other lexicalization processes involving prefixes. The assumption is that the underived forms that do not have a derivational prefix could exhibit the original onset, while the derived forms show a form with a fossilized prefix. In the broad cognate assignment, such alternations are treated as one set, while in the fine-grained assignment they are kept distinct.

An instance of this is found in the cognate sets for SHEEP and COTTON (set 157). In all three branches there is a word 'sheep' derived from 'cotton' with the animate prefix. For Mixtec, we can reconstruct PM **kati?* 'cotton' and **ti-kati?* 'sheep' [ANIM-cotton] with matching initial consonant. The same goes for Cuicatec, which has *kutʃi* 'cotton' and *iti-kutʃi* 'sheep' [ANIM-cotton]. In Triqui, however, we find *katʃi*^{h} 'cotton' but *ma*^{3}*tsi*^{h32} 'sheep' in Chicahuaxtla and Copala, with a different initial consonant, and *tʃ+a*^{3}*tʃi*^{h2} in Itunyoso, with the animate prefix but missing an initial consonant in 'cotton'. It is possible that these bilabial-initial forms reflect the original onset prior to any prefixation (Christian's suggestion, add citation). As with other alternations, we do not distinguish these in the broad cognate assignments, but only in the fine-grained one.

3.2.4 Unexpected vowel correspondences

Some forms show irregular vowel correspondences, but are otherwise clearly cognate. In the broad cognate assignment, such irregularities were ignored and all the forms assigned to the same cognate set. In the fine-grained sets they are kept distinct. We find this, for example, in set 501 for NOW/TODAY in Mixtec. Most varieties show a reflex of PM **wi*^{5}*tĩ*^{5}, but some varieties have a final [a] instead, so for example San Juan Diuxi Mixtec [βi^{5}t^{n}a^{1}]. Triqui and Cuicatec do not have reflexes of this set. Based on Figure 2, it seems that this phenomenon is characteristic of the Alta region.

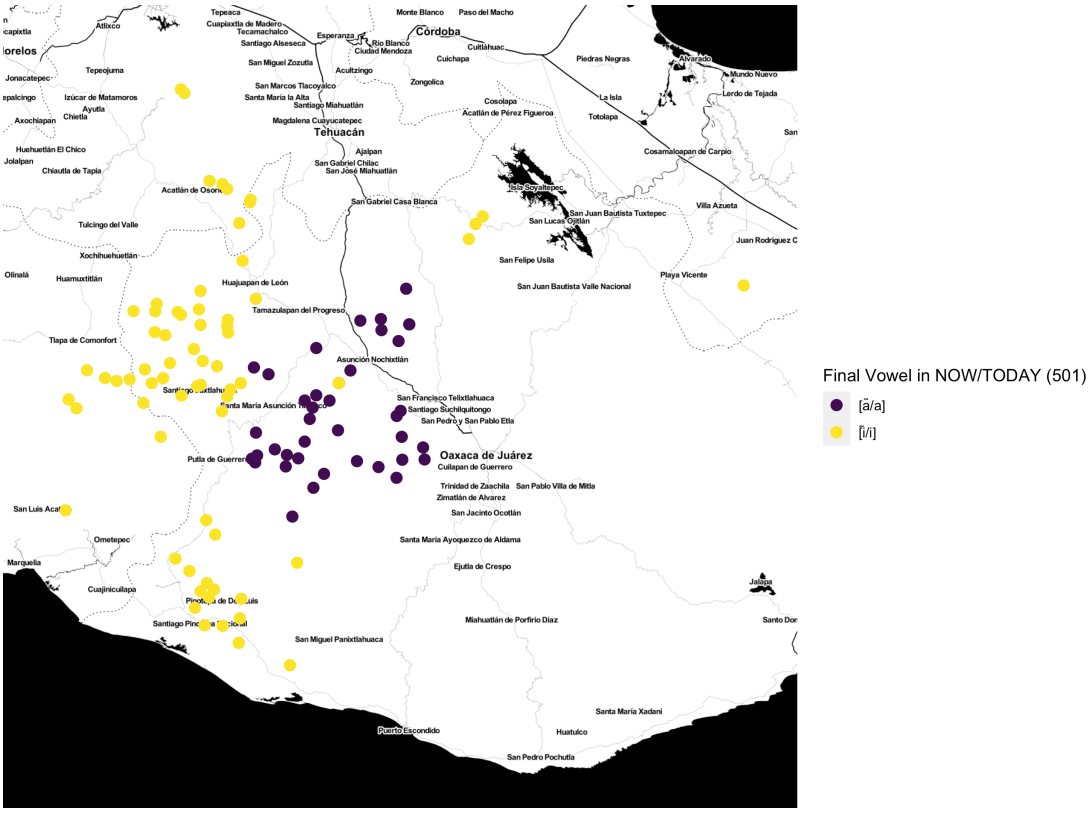

Figure 2: Final vowel of set 501 in Mixtec varieties

4 Q-Residuals and Delta-Scores per language

In this section, we list the Q-residuals and delta scores for each language for both the broad and the fine-grained cognate assessment.

Table 6: Q-residuals and δ-scores - broad cognate assignment

ID	DOCULECT	Delta Score	Q-residual
96	SantoDomingoHuendioMixtec	0.39447	0.016673
16	PinotepaDonLuisMixtec	0.35242	0.016808
59	SanPedroTidaaMixtec	0.38935	0.017583
55	SanPedroAtoyacMixtec	0.34916	0.0178
87	SantiagoIxtayutlaMixtec	0.38839	0.018122
3	AlcozaucaGuerreroMixtec	0.37964	0.018392
4	CahuatacheMixtec	0.39076	0.018439
56	SanPedroJicayanMixtec	0.35316	0.018543
78	SantaMariaPenolesMixtec	0.38371	0.018558
42	SanLorenzoMixtec	0.36563	0.018579
24	SanAntonioHuitepecMixtec	0.4084	0.018652
7	CosoltepecMixtec	0.37147	0.018699
72	SantaMariaAcatepecMixtec	0.35375	0.018751
48	SanMartinPerasMixtec	0.36788	0.018755
53	SanMiguelElGrandeMixtec	0.38087	0.018826
89	SantiagoJuxtlahuacaMixtec	0.40592	0.018927
73	SantaMariaApazcoMixtec	0.37693	0.019022
12	IxpantepecNievesMixtec	0.35578	0.019206
67	SantaCatarinaMechoacanMixtec	0.35241	0.019293
94	SantiagoTlazoyaltepecMixtec	0.37836	0.019371
40	SanJuanTeitaMixtec	0.39269	0.01958
60	SanPedroTututepecMixtec	0.36666	0.019696
27	SanBartoloSoyaltepecMixtec	0.37827	0.019859
64	SanSebastianTecomaxtlahuacaMixtec	0.38161	0.019969
102	TepejilloMixtec	0.36701	0.020008
110	ZapotitlanPalmasMixtec	0.39056	0.020058
66	SantaCatarinaEstetlaMixtec	0.39258	0.020067
86	SantiagoIxtaltepecMixtec	0.3946	0.020073
52	SanMiguelChicahuaMixtec	0.38364	0.020135
88	SantiagoJamiltepecMixtec	0.36607	0.020256
108	YucunutiBenitoJuarezMixtec	0.3488	0.020282
99	SantosReyesTepejilloMixtec	0.39613	0.020356
76	SantaMariaJicaltepecMixtec	0.35483	0.020439
49	SanMateoSindihuiMixtec	0.38757	0.020446
79	SantaMariaYolotepecMixtec	0.39413	0.020547
85	SantiagoChazumbaMixtec	0.38605	0.02055
38	SanJuanNumiMixtec	0.40653	0.020551
19	SanAgustinTlacotepecMixtec	0.39042	0.020577
15	PiedraAzulMixtec	0.37042	0.020734

ID	DOCULECT	Delta Score	Q-residual
69	SantaCruzBravoMixtec	0.39342	0.020744
95	SantiagoYosonduaMixtec	0.38334	0.020807
103	TlahuapaMixtec	0.36028	0.02097
101	TepangoMixtec	0.4121	0.021074
39	SanJuanTamazolaMixtec	0.37616	0.021187
29	SanFranciscoSayultepecMixtec	0.38279	0.02124
70	SantaCruzItundujiaMixtec	0.38438	0.021249
6	CoicoyanlasFloresMixtec	0.36507	0.021321
50	SanMiguelAchiutlaMixtec	0.40013	0.021367
71	SantaLuciaMonteverdeMixtec	0.38759	0.021518
75	SantaMariaHuazolotitlanMixtec	0.34994	0.021621
83	SantiagoApoalaMixtec	0.38061	0.021635
26	SanBartolomeYucuaneMixtec	0.40592	0.021645
93	SantiagoTilantongoMixtec	0.39057	0.02168
17	SanAgustinAtenangoMixtec	0.3387	0.021694
98	SantoDomingoTonahuixtlaMixtec	0.37046	0.021751
54	SanMiguelPiedrasMixtec	0.39582	0.021756
65	SantaAnaCuauhtemocMixtec	0.38574	0.02182
97	SantoDomingoNuxaaMixtec	0.39764	0.02191
25	SanAntonioTepetlapaMixtec	0.37313	0.021916
8	CuatzoquitengoMixtec	0.39184	0.021981
63	SanSebastianMonteMixtec	0.34536	0.022421
104	XayacatlanBravoMixtec	0.38031	0.022537
9	CuyamecalcoVillaZaragozaMixtec	0.39381	0.022688
46	SanMartinEstadoMixtec	0.34515	0.022817
91	SantiagoPinotepaNacionalMixtec	0.37041	0.022836
45	SanMartinDuraznosMixtec	0.39128	0.022871
58	SanPedroMolinosMixtec	0.39885	0.023003
90	SantiagoNuyooMixtec	0.40014	0.023009
57	SanPedroJocotipacMixtec	0.4022	0.023154
80	SantaMariaYucuhitiMixtec	0.40212	0.023168
82	SantaMariaZacatepecMixtec	0.41518	0.0232
51	SanMiguelAhuehuetitlanMixtec	0.35373	0.023213
100	SantoTomasOcotepecMixtec	0.4305	0.023751
44	SanMarcoslaFlorMixtec	0.37978	0.023792
37	SanJuanMixtepecMixtec	0.40831	0.02458
107	YucunaniMixtec	0.40457	0.025464
32	SanJorgeNuchitaMixtec	0.3512	0.025551
11	GuadalupeVillahermosaMixtec	0.34583	0.025553
81	SantaMariaYucunicocoMixtec	0.4101	0.025594
92	SantiagoTamazolaMixtec	0.3453	0.025853
74	SantaMariaChigmecatitlanMixtec	0.41898	0.026886
84	SantiagoCacaloxtepecMixtec	0.40514	0.026984
23	SanAndresYutatioMixtec	0.3841	0.027326
106	YoloxochitlMixtec	0.43969	0.027349
28	SanEstebanAtatlahucaMixtec	0.40908	0.027517
30	SanJeronimoProgresoMixtec	0.35682	0.027541

ID	DOCULECT	Delta Score	Q-residual
10	ElJicaralMixtec	0.38521	0.027881
43	SanLuisMoreliaMixtec	0.35751	0.028735
34	SanJuanColoradoMixtec	0.40052	0.02893
33	SanJuanCoatzospamMixtec	0.43118	0.02964
5	ChalcatongoHidalgoMixtec	0.4232	0.029882
22	SanAndresNuxinoMixtec	0.3897	0.030582
14	MagdalenaPenascoMixtec	0.44773	0.031618
31	SanJeronimoXayacatlanMixtec	0.4058	0.032198
13	LaBateaMixtec	0.41863	0.033276
68	SantaCatarinaTlaltempanMixtec	0.40918	0.033835
18	SanAgustinChayucoMixtec	0.42003	0.035519
36	SanJuanDiuxiMixtec	0.43193	0.035598
109	YucuquimiOcampoMixtec	0.36772	0.036917
62	SanPedroySanPabloTeposcolula1600Mixtec	0.47116	0.037287
61	SanPedroYosonamaMixtec	0.45075	0.0434
35	SanJuanCopalaTriqui	0.41682	0.044896
47	SanMartinItunyosoTriqui	0.42385	0.046283
2	AlacatlatzalaMixtec	0.46709	0.047956
20	SanAndresChicahuaxtla1890Triqui	0.42597	0.048392
105	XochapaMixtec	0.44705	0.050993
1	AbasoloValleMixtec	0.45764	0.052056
21	SanAndresChicahuaxtlaTriqui	0.43615	0.053047
77	SantaMariaPapaloCuicatec	0.43383	0.060904
41	SanJuanTepeuxila1900Cuicatec	0.43642	0.066965

Table 7: Q-residuals and δ-scores - broad cognate assignment

ID	DOCULECT	Delta Score	Q-residual
16	PinotepaDonLuisMixtec	0.34219	0.016147
96	SantoDomingoHuendioMixtec	0.37743	0.016578
59	SanPedroTidaaMixtec	0.37366	0.016597
42	SanLorenzoMixtec	0.35671	0.017271
78	SantaMariaPenolesMixtec	0.38678	0.017987
4	CahuatacheMixtec	0.37845	0.018046
64	SanSebastianTecomaxtlahuacaMixtec	0.35837	0.01806
24	SanAntonioHuitepecMixtec	0.38902	0.018104
56	SanPedroJicayanMixtec	0.34601	0.018113
73	SantaMariaApazcoMixtec	0.3639	0.018171
89	SantiagoJuxtlahuacaMixtec	0.38933	0.018183
7	CosoltepecMixtec	0.35761	0.018213
48	SanMartinPerasMixtec	0.36533	0.018313
72	SantaMariaAcatepecMixtec	0.34623	0.018345
3	AlcozaucaGuerreroMixtec	0.36654	0.018414
53	SanMiguelElGrandeMixtec	0.36098	0.018437
12	IxpantepecNievesMixtec	0.34615	0.018475
67	SantaCatarinaMechoacanMixtec	0.34677	0.018524
55	SanPedroAtoyacMixtec	0.34967	0.01874
76	SantaMariaJicaltepecMixtec	0.35071	0.018794
60	SanPedroTututepecMixtec	0.35617	0.018868
99	SantosReyesTepejilloMixtec	0.37173	0.018917
108	YucunutiBenitoJuarezMixtec	0.33067	0.01909
52	SanMiguelChicahuaMixtec	0.36184	0.019096
95	SantiagoYosonduaMixtec	0.36234	0.019131
102	TepejilloMixtec	0.35388	0.019167
65	SantaAnaCuauhtemocMixtec	0.37865	0.019243
29	SanFranciscoSayultepecMixtec	0.3679	0.019279
9	CuyamecalcoVillaZaragozaMixtec	0.37708	0.019336
27	SanBartoloSoyaltepecMixtec	0.36656	0.019363
101	TepangoMixtec	0.39948	0.019533
87	SantiagoIxtayutlaMixtec	0.38758	0.019568
86	SantiagoIxtaltepecMixtec	0.37174	0.019703
98	SantoDomingoTonahuixtlaMixtec	0.35544	0.019718
19	SanAgustinTlacotepecMixtec	0.3786	0.019764
49	SanMateoSindihuiMixtec	0.37308	0.019872
6	CoicoyanlasFloresMixtec	0.35649	0.019983
17	SanAgustinAtenangoMixtec	0.32706	0.020008
66	SantaCatarinaEstetlaMixtec	0.3866	0.02007
69	SantaCruzBravoMixtec	0.37935	0.020105
93	SantiagoTilantongoMixtec	0.38273	0.020127
39	SanJuanTamazolaMixtec	0.36577	0.020137
40	SanJuanTeitaMixtec	0.38732	0.020191
88	SantiagoJamiltepecMixtec	0.36016	0.020238

ID	DOCULECT	Delta Score	Q-residual
83	SantiagoApoalaMixtec	0.3625	0.020277
75	SantaMariaHuazolotitlanMixtec	0.3413	0.020306
110	ZapotitlanPalmasMixtec	0.37535	0.020318
104	XayacatlanBravoMixtec	0.35941	0.020379
94	SantiagoTlazoyaltepecMixtec	0.37107	0.020494
85	SantiagoChazumbaMixtec	0.37276	0.020586
54	SanMiguelPiedrasMixtec	0.37614	0.020715
70	SantaCruzItundujiaMixtec	0.36741	0.020751
50	SanMiguelAchiutlaMixtec	0.38043	0.020937
38	SanJuanNumiMixtec	0.39822	0.02095
25	SanAntonioTepetlapaMixtec	0.36415	0.021002
15	PiedraAzulMixtec	0.37165	0.021085
97	SantoDomingoNuxaaMixtec	0.37667	0.021099
71	SantaLuciaMonteverdeMixtec	0.38094	0.0214
80	SantaMariaYucuhitiMixtec	0.38189	0.021615
46	SanMartinEstadoMixtec	0.33382	0.021706
91	SantiagoPinotepaNacionalMixtec	0.36288	0.021796
26	SanBartolomeYucuaneMixtec	0.39571	0.021854
100	SantoTomasOcotepecMixtec	0.41402	0.021945
58	SanPedroMolinosMixtec	0.38153	0.021964
82	SantaMariaZacatepecMixtec	0.40433	0.022016
63	SanSebastianMonteMixtec	0.33881	0.022047
57	SanPedroJocotipacMixtec	0.38525	0.022174
51	SanMiguelAhuehuetitlanMixtec	0.34436	0.022236
79	SantaMariaYolotepecMixtec	0.38373	0.022271
8	CuatzoquitengoMixtec	0.37837	0.022309
45	SanMartinDuraznosMixtec	0.37827	0.022542
103	TlahuapaMixtec	0.36357	0.022601
92	SantiagoTamazolaMixtec	0.32749	0.022673
90	SantiagoNuyooMixtec	0.39185	0.023479
37	SanJuanMixtepecMixtec	0.40552	0.023489
44	SanMarcoslaFlorMixtec	0.37998	0.023677
28	SanEstebanAtatlahucaMixtec	0.38158	0.02371
32	SanJorgeNuchitaMixtec	0.33855	0.024182
84	SantiagoCacaloxtepecMixtec	0.38472	0.024473
11	GuadalupeVillahermosaMixtec	0.33318	0.024481
106	YoloxochitlMixtec	0.42065	0.024642
10	ElJicaralMixtec	0.37445	0.025651
81	SantaMariaYucunicocoMixtec	0.40613	0.025722
107	YucunaniMixtec	0.40427	0.026279
34	SanJuanColoradoMixtec	0.38754	0.027052
5	ChalcatongoHidalgoMixtec	0.39944	0.027079
33	SanJuanCoatzospamMixtec	0.42103	0.027125
23	SanAndresYutatioMixtec	0.36429	0.027267
74	SantaMariaChigmecatitlanMixtec	0.41314	0.027269
30	SanJeronimoProgresoMixtec	0.34841	0.027676
43	SanLuisMoreliaMixtec	0.34884	0.028022

ID	DOCULECT	Delta Score	Q-residual
14	MagdalenaPenascoMixtec	0.42358	0.028497
22	SanAndresNuxinoMixtec	0.37284	0.029256
31	SanJeronimoXayacatlanMixtec	0.38606	0.029739
68	SantaCatarinaTlaltempanMixtec	0.40679	0.032199
13	LaBateaMixtec	0.42011	0.033348
62	SanPedroySanPabloTeposcolula1600Mixtec	0.44645	0.033635
18	SanAgustinChayucoMixtec	0.41409	0.035336
36	SanJuanDiuxiMixtec	0.4187	0.036071
109	YucuquimiOcampoMixtec	0.36156	0.039198
35	SanJuanCopalaTriqui	0.40477	0.041534
105	XochapaMixtec	0.41849	0.042476
61	SanPedroYosonamaMixtec	0.42295	0.04265
2	AlacatlatzalaMixtec	0.46183	0.048779
47	SanMartinItunyosoTriqui	0.41767	0.048815
20	SanAndresChicahuaxtla1890Triqui	0.42416	0.054702
1	AbasoloValleMixtec	0.45486	0.055167
21	SanAndresChicahuaxtlaTriqui	0.4343	0.055341
77	SantaMariaPapaloCuicatec	0.42673	0.056234
41	SanJuanTepeuxila1900Cuicatec	0.43647	0.063332

References

Alexander, Ruth Mary. 1980. *Gramática mixteca de Atatlahuca.* Instituto Lingüístico de Verano, A.C.

Alexander, Ruth Mary. 1988. A syntactic sketch of Ocotepec Mixtec. In C Henry Bradley & Barbara E Hollenbach (eds.), *Studies in the syntax of Mixtecan languages*, vol. 1, 151–304. Dallas & Texas: Summer Institute of Linguistics and the University of Texas at Arlington.

de Alvarado, Francisco. 1962 [1593]. *Vocabulario en lengua mixteca. Reproducción facsimilar con un estudio de Wigberto Jiménez Moreno.* México, D.F.: Instituto Nacional Indigenista e Instituto Nacional de Antropología e Historia.

Amith, Jonathan D. & Rey Castillo García. n.d. Recursos lexicosemánticos para el mixteco de Yoloxóchitl, municipio de San Luis Acatlán, Guerrero (Glottocode yolo1241; ISO 639-3 xty). unpublished.

Anderson, E Richard & Hilario Concepción Roque. 1983. *Diccionario Cuicateco: Español-Cuicateco, Cuicateco-Español.* México, D.F.: Instituto Lingüístico de Verano.

Anderson, Lynn. 2006. *Vocabulario de palabras que se relacionan con el maíz en mixteco de Alacatlatzala, Guerrero.* México, D.F.: Instituto Lingüístico de Verano, A.C. 2nd edn.

Becerra Roldán, Braulio. 2015. *Un estudio fonológico del mixteco de Santo Domingo Huendio, Oaxaca.* México, D.F. ENAH MA thesis.

Belmar, Francisco. 1897. *Ensayo sobre la lengua trike* Lenguas Indígenas del Estado de Oaxaca. Lorenzo San-Germán.

Belmar, Francisco. 1902. *El cuicateco.* Imprenta del Comercio.

Bradley, C. Henry & J. Kathryn Josserand. 1982. El protomixteco y sus descendientes. *Anales de Antropología* 19(2). 279–343.

Bradley, David P. 1991. A preliminary syntactic sketch of Concepción Pápalo Cuicatec. In C. Henry Bradley & Barbara E. Hollenbach (eds.), *Studies in the Syntax of Mixtecan languages*, vol. 3, 409–506. Summer Institute of Linguistics and the University of Texas at Arlington.

Campbell, Lyle. 1997. *American Indian Languages: The Historical Linguistics of Native America.* Oxford: Oxford University Press.

Chao, Yuen-Ren. 1930. ə sistim əv "toun-letəz" [A system of tone letters]. *Le maître phonétique* 8(30). 24–27.

Daly, John P. 1978. Notes on Diuxi Mixtec tone. *Work Papers of the Summer Institute of Linguistics, University of North Dakota Session* 22(1).

DiCanio, Christian T. 2008. *The Phonetics and Phonology of San Martín Itunyoso Trique*: University of California, Berkeley PhD dissertation.

Dürr, Michael. 1987. A preliminary reconstruction of the Proto-Mixtec tonal system. *Indiana* 11. 19–61.

Eberhard, David M., Gary F. Simons & Charles D. Fennig. 2021. *Ethnologue: Languages of the World.* Dallas, Texas: SIL International 24th edn.

Farris, Edwin R. 1992. A syntactic sketch of Yosondúa Mixtec. In C Henry Bradley & Barbara E Hollenbach (eds.), *Studies in the syntax of Mixtecan languages*, vol. 4, 1–171. Summer Institute of Linguistics and the University of Texas at Arlington.

Galindo Sánchez, Bernardo. 2009. *Vocabulario Básico Tu'un Savi - Castellano.* Xalapa, Veracruz: Academia Veracruzana de las Lenguas Indígenas 1st edn.

Gittlen, Laura. 2016. *Gramática popular Mixteco del norte de Tlaxiaco.* Instituto Lingüístico de Verano.

Good, Claude. 1978. *Diccionario triqui de Chicahuaxtla: triqui-castellano, castellano-triqui.* México, D.F.: Instituto Lingüístico de Verano, A.C.

Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2021. Glottolog 4.4. Available online at http://glottolog.org.

Hernández Martínez, Carmen, Eric W. Campbell & Griselda Reyes Basurto. 2021. MILPA (Mexican Indigenous Language Promotion and Advocacy): A community-centered linguistic collaboration supporting Indigenous Mexican languages in California. In Justyna Olko & Julia Sallabank (eds.), *Revitalizing Endangered Languages: A Practical Guide*, 216–217. Cambridge University Press.

Hernández Mendoza, Fidel. 2020. Vocabulario Triqui de Chicahuaxtla. unpublished.

Hills, Robert A. 1990. A syntactic sketch of Ayutla Mixtec. In C. Henry Bradley & Barbara E. Hollenbach (eds.), *Studies in the syntax of Mixtecan languages*, vol. 2, Summer Institute of Linguistics and the University of Texas at Arlington.

Hinton, Leanne, Gene Buckley, Marv Kramer & Michael Meacham. 1991. Preliminary analysis of Chalcatongo Mixtec tone. In *Papers from the American Indian Languages Conference, University of California, Santa Cruz, July and August*, 147–155.

Hollenbach, Barbara E. 1992. A syntactic sketch of Copala Trique. In C. Henry Bradley & Barbara E. Hollenbach (eds.), *Studies in the Syntax of Mixtecan languages*, Summer Institute of Linguistics.

Hollenbach, Barbara Elena Erickson. 2013. *Gramática del mixteco de Magdalena Peñasco (Sa'an Ñuu Savi)*. Tlalpan & Mexico: Instituto Lingüístico de Verano, A.C.

Hollenbach, Barbarba Elena Erickson. 2017. *Diccionario mixteco de Magdalena Peñasco*. Instituto Lingüístico de Verano, A.C.

Johnson, Audrey F. 1988. A syntactic sketch of Jamiltepec Mixtec. In C Henry Bradley & Barbara E Hollenbach (eds.), *Studies in the Syntax of Mixtecan languages*, vol. 1, 11–150. Summer Institute of Linguistics and the University of Texas at Arlington.

Josserand, Judy Kathryn. 1983. *Mixtec dialect history*: University of Michigan Ann Arbor PhD dissertation.

Kiddle, Lawrence B. 1964. American Indian reflexes of two Spanish words for cat. *International Journal of American Linguistics* 30(3). 299–305.

Kuiper, Albertha & Joy Oram. 1991. A syntactic sketch of Diuxi-Tilantongo Mixtec. In C Henry Bradley & Barbara E Hollenbach (eds.), *Studies in the syntax of Mixtecan languages*, vol. 3, 185–408. Summer Institute of Linguistics and the University of Texas at Arlington.

Longacre, Robert E. 1957. *Proto-Mixtecan*. Indiana University.

Longacre, Robert E. 1961. Swadesh's Macro-Mixtecan Hypothesis. *International Journal of American Linguistics* 27(1). 9–29.

Macaulay, Monica Ann. 1996. *A grammar of Chalcatongo Mixtec*. University of California Press.

Mak, Cornelia & Robert Longacre. 1960. Proto-Mixtec phonology. *International Journal of American Linguistics* 26(1). 23–40.

Marlett, Stephen A. 1992. Nasalization in Mixtec languages. *International Journal of American Linguistics* 58(4). 425–435.

Matsukawa, Kosuke. 2005. *Preliminary Reconstruction of Proto-Triqui*. Albany NY State University of New York at Albany MA thesis.

Mendoza, Inî G. & Simon L. Peters. 2020. Vocabulario del Tù'un Sàjvĭ (Mixteco) de Piedra Azul y Paredón, San Martín Peras. unpublished.

North, Joanne & Jäna Shields. 1977. Silacayoapan mixtec phonology. In William R. Merrifield (ed.), *Studies in Otomanguean Phonology*, 21–33. Summer Institute of Linguistics and the University of Texas at Arlington.

Ohala, John J. 1983. The origin of sound patterns in vocal tract constraints. In *The production of speech*, 189–216. Springer.

Ohala, John J. 1989. Sound change is drawn from a pool of synchronic variation. In *Language change*, 173–198. De Gruyter Mouton.

Okura, Eve. 2015. Language vs. dialect in language cataloguing: The vexed case of Otomanguean dialect continua. *University of Hawai 'i at Mānoa Working Papers in Linguistics* 46(5). 1–19.

Pensinger, Brenda J et al. 1974. *Diccionario Mixteco: mixteco del este de Jamiltepec, pueblo de Chayuco*. México, D.F.: Instituto Lingüístico de Verano, A.C.

Pike, Eunice V & Joy Oram. 1976. Stress and tone in the phonology of Diuxi Mixtec. *Phonetica* 33(5). 321–333.

Reyes Basurto, Griselda. 2020. Diccionario del Mixteco de Tlahuapa, Guerrero, Mexico. unpublished.

Schultze-Jena, Leonhard. 1938. *Bei den Azteken, Mixteken und Tlapaneken der Sierra Madre del Sur von*

Mexico, vol. 3 Indiana. Jena: Gustav Fischer.

Shields, Jana. 1988. A syntactic sketch of Silacayoapan Mixtec. In C. Henry Bradley & Barbara E. Hollenbach (eds.), *Studies in the Syntax of Mixtecan Languages*, vol. 1, 305–449. Summer Institute of Linguistics and the University of Texas at Arlington.

Small, Priscilla C. 1990. A syntactic sketch of Coatzospan Mixtec. In *Studies in the Syntax of Mixtecan languages*, vol. 2, 261–479. Summer Institute of Linguistics and the University of Texas at Arlington.

Solano, Juvenal. 2020. San Sebastián del Monte Mixtec Vocabulary. In *Mixteco Oxnard Archivos*, unpublished.

Stark, Sharon C., Audrey P. Johnson & Benita González de Guzmán. 2013. *Diccionario básico del mixteco de Xochapa, Guerrero*. Instituto Lingüístico de Verano, A.C. 3rd edn.

Stark, Sharon C., Andrea Johnson Peterson & Filiberto Lorenzo Cruz. 1986. *Diccionario Mixteco de San Juan Colorado*. México, D.F.: Instituto Lingüístico de Verano, A.C.

Swadesh, Morris. 1960. The Oto-Manguean hypothesis and Macro Mixtecan. *International Journal of American Linguistics* 26(2). 79–111.

Swanton, Michael & Juana Mendoza Ruíz. forthcoming. Observaciones sobre la diacronía del tono en el Tu'un Savi (mixteco) de Alcozauca de Guerrero. *SLI* .

Towne, Douglas. 2011. *Gramática popular del tacuate (mixteco) de Santa María Zacatepec, Oaxaca*. Instituto Lingüístico de Verano, A.C.

Ferguson de Williams, Judith. 2007. *Gramática popular del mixteco del municipio de Tezoatlán, San Andrés Yutatío, Oaxaca*. Instituto Lingüistico de Verano, AC.

Williams, Judith F., Gerardo Ojeda Morales & Liborio Torres Benavides. 2017. *Diccionario mixteco de San Andrés Yutatío, Tezoatlán, Oaxaca*. Ciudad de México: Instituto Lingüístico de Verano, A.C.

Zylstra, Carol F. 1991. A syntactic sketch of Alacatlatzala Mixtec. In C Henry Bradley & Barbara E Hollenbach (eds.), *Studies in the Syntax of Mixtecan languages*, vol. 3, 1–177. Summer Institute of Linguistics and the University of Texas at Arlington.

Zylstra, Carol F. 2012. *Gramática del Tu'un Savi (la lengua mixteca) de Alacatlatzala, Guerrero*. Instituto Lingüístico de Verano, A.C.