Journal of Open Humanities Data

A Mixtec Sound Change Database

RESEARCH PAPER

SANDRA AUDERSET
ERIC W. CAMPBELL
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

We present an interlinked, expandable database of segmental sound changes among a large sample of Mixtec languages of Mexico. The database provides an up-to-date repository for scholars working on Mixtec and related languages. It serves the wider historical linguistics community by providing a model for managing large data sets in a way that streamlines traditional historical linguistic analysis at the same time as preparing the data for computational and quantitative analysis. We build upon previous studies but also introduce a novel annotation scheme for analyzing sound change in large and complex language groups or families. The database is hosted on GitHub (https://github.com/SAuderset/MixteCaSo) so that it can be improved and expanded on in the future. Publication versions such as this one are archived on Zenodo (https://doi.org/10.5281/zenodo.10630996).

CORRESPONDING AUTHOR:
Sandra Auderset
Department of Linguistics, University of Bern, CH
sandra.auderset@unibe.ch

KEYWORDS:
sound change; comparative method; Mixtec languages; subgrouping

TO CITE THIS ARTICLE:
Auderset, S., & Campbell, E. W. (2024). A Mixtec Sound Change Database. *Journal of Open Humanities Data*, 10: 24, pp. 1–13. DOI: https://doi.org/10.5334/johd.184

DOI: 10.5334/johd.184

1 CONTEXT AND MOTIVATION

Understanding the history of a language family enriches our accounts of the synchronic linguistic variation we observe, of the relationships between languages, and our general knowledge of processes of language change. The comparative method of historical linguistics, based on the principle of regular sound change, remains the primary tool for establishing language families and subgrouping, reconstructing proto-languages, and describing sound changes. This is true for both traditional methods, in which cognate sets, reconstructions, and family trees are established by hand, as well as computational approaches, in which part of this work is carried out by models and algorithms (Bowern, 2018; Greenhill, Heggarty, and Gray, 2020; List, Walworth, Greenhill, Tresoldi, and Forkel, 2018). Both approaches crucially rely on good data that follows the FAIR principles of data management (Wilkinson et al., 2016), so that other researchers can make use of it. With regards to historical linguistics, such databases have become available mostly in the form of word lists that can be used to determine cognacy and for applying Bayesian phylogenetic models (see for example the Lexibank project by List et al., 2022) and collections of reconstructed and attested phoneme inventories (such as the BDPROTO database by Moran, Grossman, and Verkerk, 2021).

Here, we introduce a comprehensive database of Mixtec segmental sound changes with proto-Mixtec lexical reconstructions. The database achieves several related goals: i) creating a reusable digital record of the most up-to-date comparative Mixtec lexical data, ii) systematizing and standardizing earlier materials to a common representation (IPA), and iii) reviewing and updating reconstructions and proposed sound changes in light of newly available data and insights. The database consists of modules that are linked together via common, unique identifiers. These modules cover metadata of the language sample, bibliographic information, annotated cognate sets, reconstructed proto-Mixtec forms, and sound changes. The modular approach ensures that the database can be used for a range of research questions and serves as a model for sound change databases of other language families.

Mixtec (or Tu'un Savi) refers to the languages spoken traditionally and currently by the Ñuu Savi people in southern Mexico (Julián Caballero, 1999), and in diaspora communities in other parts of Mexico and the United States. Mixtec consists of perhaps some 200 distinct local varieties,[1] many of which are not mutually intelligible. Mixtec is most closely related to the Cuicatec and Triqui languages, and these three groups together comprise the larger Mixtecan language family (Longacre, 1957, 1961), which in turn is considered one of the major branches of the highly diverse and widely spread Otomanguean stock of Mesoamerica (Campbell, 2017b; Kaufman, 2006; Rensch, 1976). The complex history of Mixtec peoples is still insufficiently understood, even though they have been and still are a large and influential group in Mesoamerica. The great number of distinct varieties and their complex relationships pose challenges not just for subgrouping, but also for the identification and analysis of sound changes, such that "splits and mergers of phonological history have somehow managed to create several major branches of Mixtec which mostly look alike despite their checkered phonological histories" (Josserand, 1983, 459).

Recent years have seen an uptake in the documentation and description of Mixtec varieties, some previously studied, and others not. This allows us to incorporate a larger and more representative sample than earlier studies to inform proto-Mixtec reconstruction and sound changes. These data and methods are of interest to scholars of Otomanguean and Mesoamerican languages, as well as historical linguists who are interested in applying computer-assisted workflows for qualitative and quantitative analysis.

2 DATASET DESCRIPTION

REPOSITORY LOCATION

https://doi.org/10.5281/zenodo.10630996 and https://github.com/SAuderset/MixteCaSo

1 We use the terms 'variety' and 'language' interchangeably, but refrain from using the term 'dialect' since it carries a negative connotation in Mexico and has been part of a long history of oppression of communities that speak Mixtecan and other indigenous languages (Cruz & Woodbury, 2014).

REPOSITORY NAME

MixteCaSo

OBJECT NAME

SAuderset/MixteCaSo-1.1.0.zip

FORMAT NAMES AND VERSIONS

tsv, PDF, R, Rmd

CREATION DATES

2022-01-01 to 2023-11-16

DATASET CREATORS

Sandra Auderset (creator, annotator), Eric W. Campbell (annotator)

LANGUAGE

English

LICENSE

CC-BY-SA-4.0

PUBLICATION DATE

2023-11-16

The data base (version 1.1.0) is available on Zenodo at https://doi.org/10.5281/zenodo.10630996. This repository also contains a document explaining the conversion from orthography to IPA and other issues regarding the source materials in more detail. The code used to produce the plots is provided as an Rmarkdown and PDF file in the same repository. The working version of the database, which is continuously updated, is on GitHub at https://github.com/SAuderset/MixteCaSo.

3 DATA COLLECTION AND STANDARDIZATION

The lexical data used as the basis for identifying the sound changes come from a recent study on subgrouping within the Mixtecan language family (Auderset et al., 2023a, 2023b). These data were originally collected through a list of 209 concepts of basic vocabulary. These entries were then cognate-coded based on the comparative method, building on previous work and our own knowledge of the languages in question. We annotated cognate morphemes (i.e. 'partial cognacy') within lexical items, as Mixtecan languages have multimorphemic words in their basic vocabulary, following List, Greenhill, and Gray (2017). More details and illustrative examples can be found in Auderset et al. (2023a, 5–6). From this data set, we selected all the Mixtec entries, setting aside the Cuicatec and Triqui data. Our sample includes 105 Mixtec varieties; an overview of all languages and sources can be found in the metadata file (data/metadata.tsv). Figure 1 shows their location as well as subgroup membership according to Auderset et al. (2023a).

The orthographies used in materials on Mixtec languages vary greatly; often each author and each source uses a system differing from all others in certain aspects. Database entries were initially collected in the source orthography and then converted to a standardized IPA representation using carefully created orthography profiles (one for each source). Some graphemes leave no ambiguity as to the sound they represent, making IPA conversion straightforward. This is the case, for example, for nasal consonants and most vowels. Other practical or linguistic graphemes, however, are ambiguous due to differences in the sound systems of the languages. Here we briefly elaborate on general conversion principles. The details regarding each variety are laid out in a supplementary file (definitions/orthography_to_ipa.pdf). Sequences of identical vowels are represented as such and not as a vowel with a length diacritic because i) the tone-bearing unit in

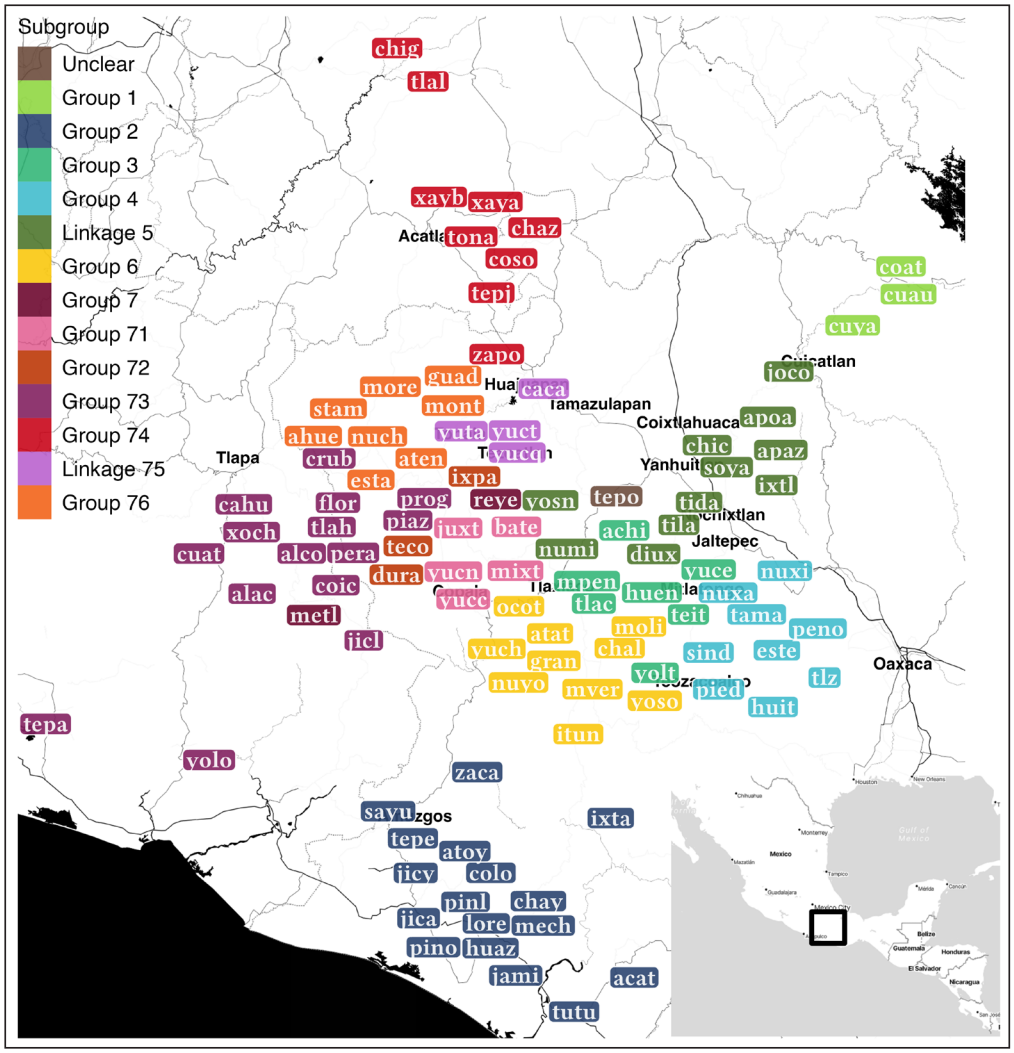

Figure 1 Map overview of sampled varieties colored by subgroup (Auderset et al., 2023a) with an inset showing the location of the detailed map within Mesoamerica.

Mixtec is analyzed as the mora (Castillo García 2007; McKendry 2013 among others), so writing each vowel separately facilitates the representation of tone, and ii) there is variation between and within Mixtec varieties in words of the form $CV_iʔV_i$, such that these can reduce to CV_iV_i. For comparative purposes, alignment across such variation is simpler if long vowels are not written as one segment. The notational conventions used to represent environments of changes are summarized in a supplementary file (definitions/orthography_to_ipa.pdf, section 2.4).

Cognate coding was carried out based on the comparative method informed by previous reconstructions, especially Josserand (1983), but also Dürr (1987), Swanton and Mendoza Ruíz (2021) and Swanton (2021). These studies incorporate and improve upon earlier reconstructions, such as Longacre's (1957) proto-Mixtecan reconstruction (based on just one variety each of Cuicatec and Triqui and four Mixtec varieties) and Mak and Longacre's (1960) proto-Mixtec reconstruction (covering 28 varieties). Notably, in these early works, only final syllables were reconstructed. Bradley and Josserand (1982) reconstructed 45 proto-Mixtec forms including initial and final syllables. They identify sound changes, relative chronologies, and present isogloss maps that point to possible contact across areas and paths of migration. Their reconstructed phoneme inventory for proto-Mixtec differs from Mak and Longacre (1960) in important ways (see Campbell, 2017a, 10). Josserand's (1983) subsequent work became state-of-the-art in Mixtec reconstruction and remains so in many respects. Focusing on vowel correspondences and changes, she compared data from 120 Mixtec varieties, reconstructed 188 proto-Mixtec forms, and classified Mixtec languages into 12 'dialect areas', some with subdivisions. Recently, Swanton (2021) provides an important reinterpretation of the Colonial Era variety of Teposcolula, including comparisons with contemporary varieties and proto-Mixtec reconstructions. Finally, Kaufman (in press) revisited Longacre's (1957) proto-Mixtecan in light of evidence from Amuzgo, Mixtecan's closest relation, and other Otomanguean branches, and he agrees with Josserand (1983) in almost every detail of the proto-Mixtec segmental sound system. As Mixtec languages exhibit multi-morphemic words in their basic vocabulary, we identified cognate morphemes within each lexical item (i.e. 'partial cognacy' sensu List et al., 2017).

	DENTAL	VELAR	LABIO-VELAR	GLOTTAL	FRONT	CENTRAL	BACK
plosive	*t*	*k*	*k*^{w}	*ʔ*	close *i*	*ɨ*	*u*
prenasalized pl.	*ⁿd*				mid *e*		*o*
nasal	*n*				open	*a*	
fricative	*s*						
affricate	*tʃ*						
approximant	*(l)*	*j*	*w*		suprasegmentals:	vowel nasality; tone	

Table 1 Proto-Mixtec phoneme inventory.

4 ISSUES IN PROTO-MIXTEC RECONSTRUCTION

The proto-Mixtec phoneme inventory that serves as the basis for the reconstruction of forms and the identification of sound changes in the database is summarized in Table 1. It is not identical to any previous source but closely resembles that of Kaufman (in press). A detailed discussion of earlier proposals with respect to our inventory is outside the scope of this paper, instead, we give a brief introduction of the most pertinent and debated issues.

The synchronic phonological representation of laryngealization/glottal stop in Mixtec languages has received considerable attention and divergent analyses. For example, Castillo García (2007) (on Yoloxóchitl), McKendry (2013) (on Nochixtlán) and Hinton et al., (1991) (on Chalcatongo) treat laryngealization as a vocalic feature, as does Gerfen (1996) for Coatzospam Mixtec, where it is automatically inserted word-medially. Macaulay and Salmons (1995) treat laryngealization as a contrastive floating feature of the root in Chalcatongo Mixtec, and Carroll (2015) and Mendoza Ruíz (2016) adopt similar analyses for Ixpantepec Nieves and Alcozauca Mixtec, respectively. North and Shields (1977) and Pike and Cowan (1967) consider it to be a glottal stop consonant in Silacayoapan and Huajuapan Mixtec, respectively. Josserand (1983) analyzes glottalization as a feature of the vowel because it was treated this way by the majority of descriptive studies available at the time. We follow Kaufman (in press) in that we include the glottal stop as a consonant, but this has practical rather than theoretical motivations. The representation of the glottal stop in the database (either as a consonant or as glottalization) does not affect the reconstruction or the characterization of sound changes in the database. In other words, the diachronic behavior of the glottal stop is such that the practical results would be no different if we analyzed it as a vowel feature. Since it is representationally simpler to write a full glottal stop, this is what is implemented in our database.[2]

We follow Swanton (2021) in reconstructing a proto-Mixtec affricate **tʃ* where older sources, such as Josserand (1983) and Mak and Longacre (1960), had a dorsal fricative **x*. Swanton (2021, 14) argues that the modern reflexes and resulting changes are easier to explain and more intuitive if one reconstructs the affricate.

We bracket proto-Mixtec **l* because it is reconstructed by Josserand (1983) and others, but almost all forms with this proto-phoneme have modern reflexes which alternate between *l* and *s* or more rarely between *l* and *ⁿd*. The alternation is not predictable and does not apply consistently across the lexicon of any one variety. It thus cannot be characterized as involving regular sound change. According to Kaufman (in press), Amuzgoan, the language group most closely related to Mixtecan, displays an alternation of nominal prefixes *ts-* for singular and *l* for plural or collective, and we interpret the unpredictable alternation between Mixtec *s* and *l* as the residue of these prefixes. We reconstruct **l* only in the cognate set 674 SMALL.SINGULAR in which modern reflexes show *l* exclusively. In all others, the reflexes with *l* are the minority, often by far, while the majority of varieties show reflexes consistent with proto-Mixtec **s* or **ⁿd*. Future analyses might show that the **l* in SMALL comes from one of these phonemes as well.

For each cognate set with sufficient data, we reconstruct a proto-Mixtec form. We re-evaluated previous reconstructions in light of the data currently available. For a few proto-forms, we propose adjustments, mostly with respect to the vowels **e* and **u*, whose diachrony is difficult to distinguish from **a* and **o*, respectively. The previous reconstructions from Dürr (1987), Josserand (1983), Swanton and Mendoza Ruíz (2021) and Swanton (2021) are provided in the

2 We note that this can easily be changed in a future/derived version of the database.

DOI: 10.5334/johd.184

database for detailed comparison. For some concepts, it is not (yet) possible to arrive at a plausible proto-Mixtec form, so not all entries were used for analyzing the sound changes. In total, the data sheet currently contains 247 lexical reconstructions and 9 classifier-like morphemes. Of the lexical protoforms, 80 (or about a third) are newly reconstructed and do not appear in previous sources.

5 CODING OF SOUND CHANGES AND STRUCTURE OF THE DATABASE

After reviewing, revising, and adding new proto-Mixtec forms for each cognate set where this is possible, we identified all regular segmental sound changes in each language of our data set. Given the large number of languages and data points, we handled this by creating multiple, interlinked databases following AUTOTYP principles such as modularity, autotypology, separation of definition and data files, and late aggregation (Witzlack-Makarevich et al., 2022). AUTOTYP is a typological database that has been continuously developed over the past twenty-five years as part of a large-scale research program to address problems that have arisen from the creation of more traditional typological databases. One of these issues is the use of fixed, *a priori* categories determined by theoretical considerations, or simply by traditional usage, which often fail to adequately capture a phenomenon across a large and diverse language sample. On the more practical side, databases are often not constructed in a way that facilitates their later re-use and expansion. This framework thus seeks to address these issues by providing guidelines and design principles for the creation of data-driven, transparent, and reusable databases. One of these principles is the use of autotypology (Bickel, 2010; Bickel et al., 2011; Bickel & Nichols, 2002), which is a typological method that does not rely on pre-defined categories, but rather on building up categories during data entry. Every time a new language is added, the existing categories are re-evaluated and expanded or modified as needed. This avoids excluding languages because they do not fit preconceived notions of a category and therefore allows the database to be largely independent of specific theoretical frameworks. The framework aims at high precision of the terms used by breaking down descriptive notions until they are unambiguous. Another important design principle is the separation of information across multiple files which are linked together via a common, standardized identifier. This flexibility makes it possible to address an array of different questions with one data set. While creating databases in this framework is initially more time-consuming than working with pre-defined categories, it provides data accuracy to a degree that is impossible with the latter (Bickel, 2007, 246). Although our study deals with only one language family and with sound changes rather than synchronic structural features, it shares multiple key components with large-scale typological studies in autotypology. First, our sample size of 105 languages is comparable to that of mid-sized typological studies (Bakker, 2010). Second, we work with a large amount of data points, rather than selecting supposedly representative examples. Finally, we build our inventory of sound changes in a bottom-up fashion, adding new changes as seen in our data set. In the following, we describe database creation and sound change coding in more detail. An overview of the full interlinked database is provided in Figure 2.

The metadata file provides information about the sample languages. This includes the language name in a standardized format which we also use as an identifier for linking across databases, the village name in its most commonly used spelling in Mexico, geographic coordinates of the village, and subgroup membership according to two previous studies (Auderset et al., 2023a; Josserand, 1983), ISO-639-3-codes and Glottocodes (where available), and the source(s) of the data. The full bibliographic information of the sources is provided in a separate bibliography file.

The cognate set database contains a unique identifier for each form, the language identifier, cognate set ID, the full form in IPA, and gloss. There are currently 15,127 such cognate coded lexical entries in the database. It can be linked to the proto-forms via the cognate IDs and to the metadata via the language identifier. An excerpt is provided in Table 2.

The protoforms file contains a unique cognate set ID, the reconstructed meaning of the cognate set, the reconstructed form in IPA, the reconstructed proto-form of earlier sources (also standardized to IPA for better comparability), and the cognate set IDs of those sources (if applicable). This database is linked with the cognate sets via the cognate set ID. An excerpt is provided in Table 3.[3]

3 PMX = Proto-Mixtec, Josserand = Josserand 1983, Durr = Dürr 1987.

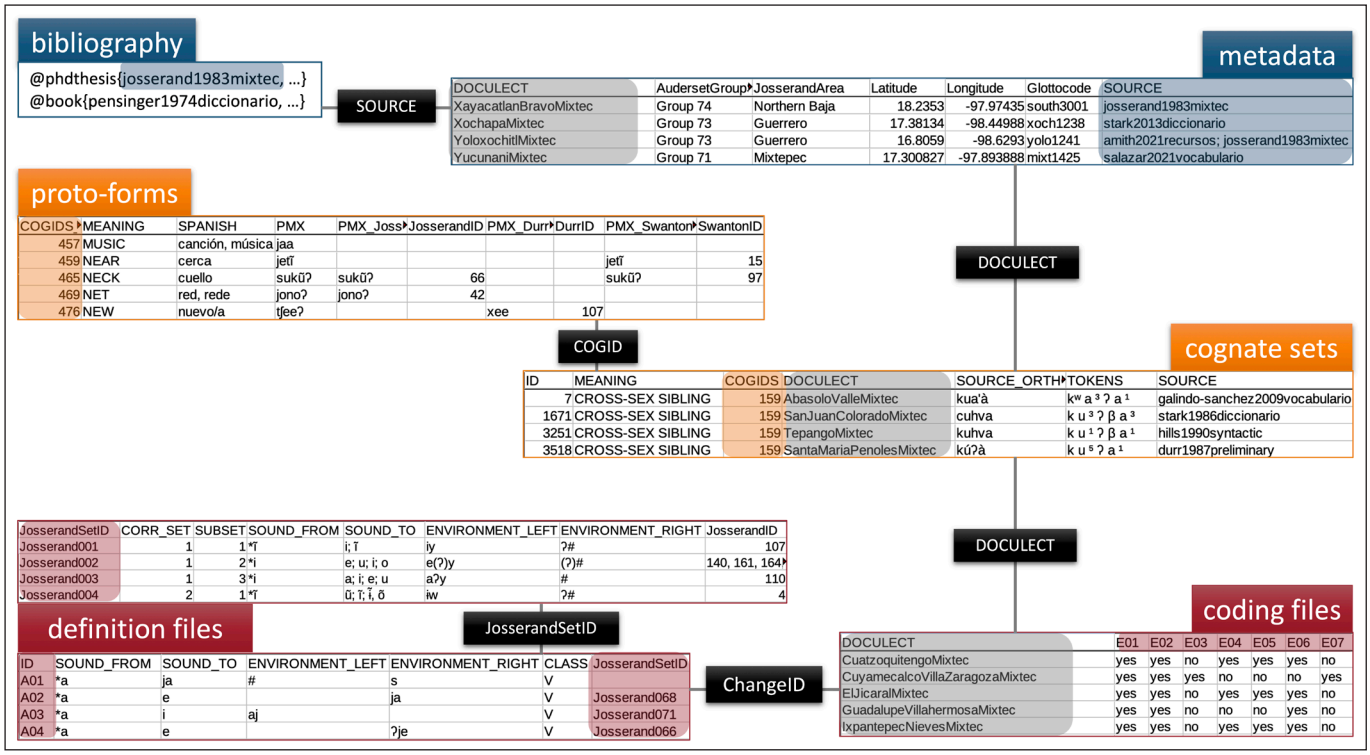

Figure 2 Schematic overview of the interlinking of files of the database. Metadata components are given in blue, source data in orange, and analysis files in red. Black boxes with white text represent the unique identifiers used to link one file with another. Possible links are shown with lines between elements.

The sound changes are distributed over a definition file containing a sound change ID, the proto-Mixtec source phoneme, the modern reflex, and the environments (split into left- and righthand) that conditioned the change, as shown in Table 4. Vowel changes also have a reference – where applicable – to the ID of the respective correspondence set from Josserand (1983). The presence or absence of each sound change in each variety was recorded in the coding file. This file contains the language identifier, sound change identifier, and value. An excerpt of the coding file is shown in Table 5. It can be linked to the definition file via the sound change ID and to the cognate sets and metadata files via the language ID. Tables 4 and 5 together exemplify the sound change coding based on the data provided in Table 2.

The presence of changes across varieties with partially overlapping targets and/or conditioning environments necessitated a very fine-grained level of coding instead of maximally generalizing changes as is typically done in work on fewer languages or in outlining the historical phonology

ID	MEANING	DOCULECT	SOURCE_ ORTHOGRAPHIC	TOKENS	COGID
2105	SALT	SanJuanDiuxiMixtec	*ñíí*	ɲ i ⁵ i ⁵	624
20187	SALT	SanMartinDuraznosMixtec	*ñìì*	ɲ i ¹ i ¹	624
6646	SALT	SantaMariaZacatepecMixtec	*ñii*	ɲ i i	624
2127	RIVER	SanJuanDiuxiMixtec	*yúté*	ʒ u ⁵ t e ⁵	607
19559	RIVER	SanMartinDuraznosMixtec	*yìtxa*	ʒ i ¹ tɕ a ³	607
12814	RIVER	SantaMariaZacatepecMixtec	*ju²ɟa²¹*	j u ³ tʃ a ³¹	607

Table 2 Excerpt from the cognate database.

COGID	MEANING	PMX	PMX_ JOSSERAND	JOSSERANDID	PMX_ DURR	DURRID
294	GRASS	**ite*				
662	SLOW	**kʷeji*	kʷeje	163		
607	RIVER	**jute*	jute	23	jute	58
107	CHILI PEPPER	**jaʔaʔ*			jaʔaʔ	12
624	SALT	**jiĭʔ*	jiĭʔ	41		

Table 3 Excerpt from the proto-forms database.[3]

of a single language. If the data were lacking or inconclusive for a particular change in a language, this is indicated with NA (not applicable).

ID	SOUND_FROM	SOUND_TO	ENVIRONMENT_LEFT	ENVIRONMENT_RIGHT
E17	*e	a	{i,u}t	
U32	*u	i	j	te
Y01	*ɨ	i		(ʔ)#
J03	*j	ɲ	#	V(ʔ)Ṽ
T12	*t	tɕ		{ɨ,e}
T14	*t	tʃ		e

Table 4 Sound change variables derived from the data in Table 2 and the protoforms in Table 3.

DOCULECT	E17	U32	Y01	J03	T12	T14
SanJuanDiuxiMixtec	no	no	no	yes	no	no
SanMartinDuraznosMixtec	yes	yes	yes	yes	yes	no
SantaMariaZacatepecMixtec	yes	no	yes	yes	no	yes

Table 5 Sound change coding based on the variables in Table 4.

We established the sound changes by evaluating each cognate set for which we have a reconstructed proto-Mixtec form in light of the modern reflexes using the comparative method. We did not code for fine phonetic detail and we excluded cases of more sporadic changes such as sibilant harmony, after confirming the accuracy of the relevant reconstructions. To allow the database to be expanded upon with more data in future work, we largely refrained from specifying environments with sound classes and rather listed conditioning environments separately. This no doubt has led to under-generalization for some changes, but such generalizations can be recovered through later aggregation (see Section 5). Additionally, the method of perhaps over-specifying the details of changes allows for the identification of 'nested', or partially-shared changes that may reflect how conditioning environments have evolved over time in subsets of varieties. In the future, this will allow us to code the relative ordering of these changes within each language and groups of languages and thus address a limitation of the current version of the database. We identified 252 sound changes, of which 133 pertain to vowels and 119 to consonants.

6 USE CASES

There are many imaginable uses of the database, from investigating subgrouping at various levels to tracing the development of specific sounds. Here we briefly discuss a few possibilities of summarizing and visualizing the data for select research questions. We exclude 29 sound changes from these analyses due to low coverage, defined here as being coded for less than two-thirds of the sample languages.

SUMMARIZING REFLEXES OF A PROTO-SOUND AND CONDITIONING ENVIRONMENTS

The sound change database can be used to summarize reflexes of a given proto-sound across all varieties and within proposed subgroups. We exemplify this with proto-Mixtec *s. Modern reflexes of this sound include no change (i.e. retention of /s/), loss, and fricatives /ð/, /ʃ/, and /h/. Three varieties (Alacatlatzala, Cahuatache, and San Luis Morelia Mixtec) retain /s/ unchanged in all environments. This is not explicitly coded in the database – as a retention is equivalent to the absence of a change. However, aggregating over all change variables related to PM *s and extracting the varieties that have no changes at all gives us exactly this information. In addition to the modern reflexes, we also want to know the conditioning environments of the changes across the sample. This information can be directly extracted and aggregated from the definition file. Figure 3 summarizes all this information as a map display. Displays like these are useful for investigating geographical patterning. We can see, for example, that /ð/ reflexes

DOI: 10.5334/johd.184

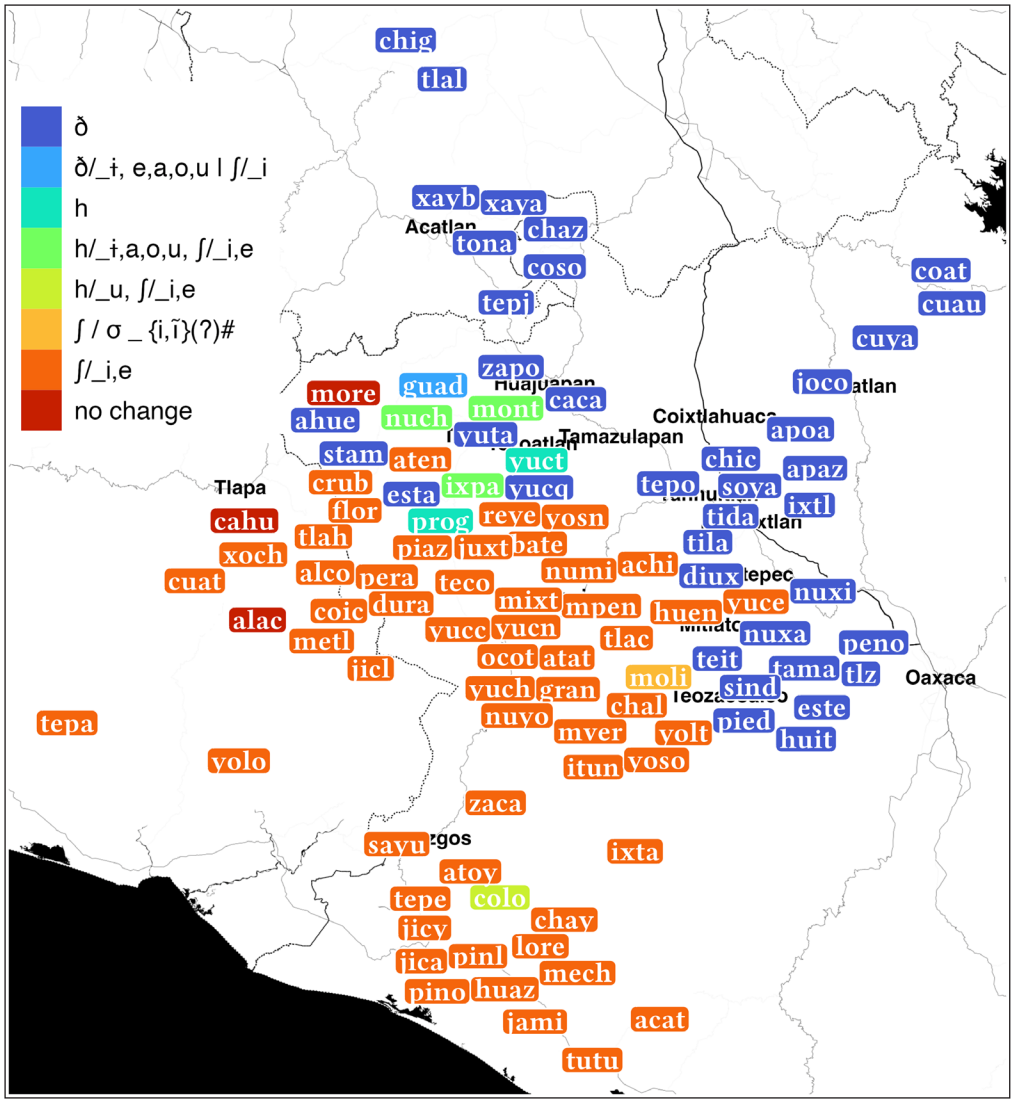

Figure 3 Distribution of primary reflexes and their conditioning environments of changes to proto-Mixtec **s*. If no environment is given, this means the change is global.

are restricted to the north and east. We can also observe that the three varieties which retain /s/ throughout are found at the western edge of the Mixtec region.

LOOKING AT SPECIFIC TYPES OF CHANGES

The database can also be used to look at specific types of changes, such as the loss of a sound or palatalization of different types of stops. We illustrate this here with palatalization of PM **t* before *i*, represented in our database as variable T01. The presence and absence of this change and the distribution across subgroups is summarized in Figure 4. Palatalizations as represented by this change variable are very common in the languages of the world so its presence in most Mixtec varieties could imaginably be due to parallel innovation. However, it is absent precisely in the two groups (Group 2 on the coast and Group 1 in the far north-east of the Mixteca) that likely represent early migrations, i.e. early split-off events (Auderset et al., 2023a; Bradley & Josserand, 1982). This distribution suggests that the change took place after Groups 1 and 2 split from the rest of Mixtec.

SUMMARIZING CHANGES ACROSS VARIETIES

As a final example, we can summarize the number of changes across varieties and look into different classes of changes (e.g. vowels vs. consonants). The varieties with very few changes overall are phonologically conservative, while those reflecting a greater number of changes can be thought of as innovative. Figure 5 illustrates this by showing the percentage of changes reflected in each variety on a map. Overall, the most conservative varieties are found in the far north-east (Auderset et al. (2023a)'s Group 1) and on the coast (Group 2), as well as scattered in the south-eastern part of the Mixteca (also known as the Eastern Alta). The most innovative varieties are clustered together in the north in an area known as the Central Baja and Tezoatlán.

DOI: 10.5334/johd.184

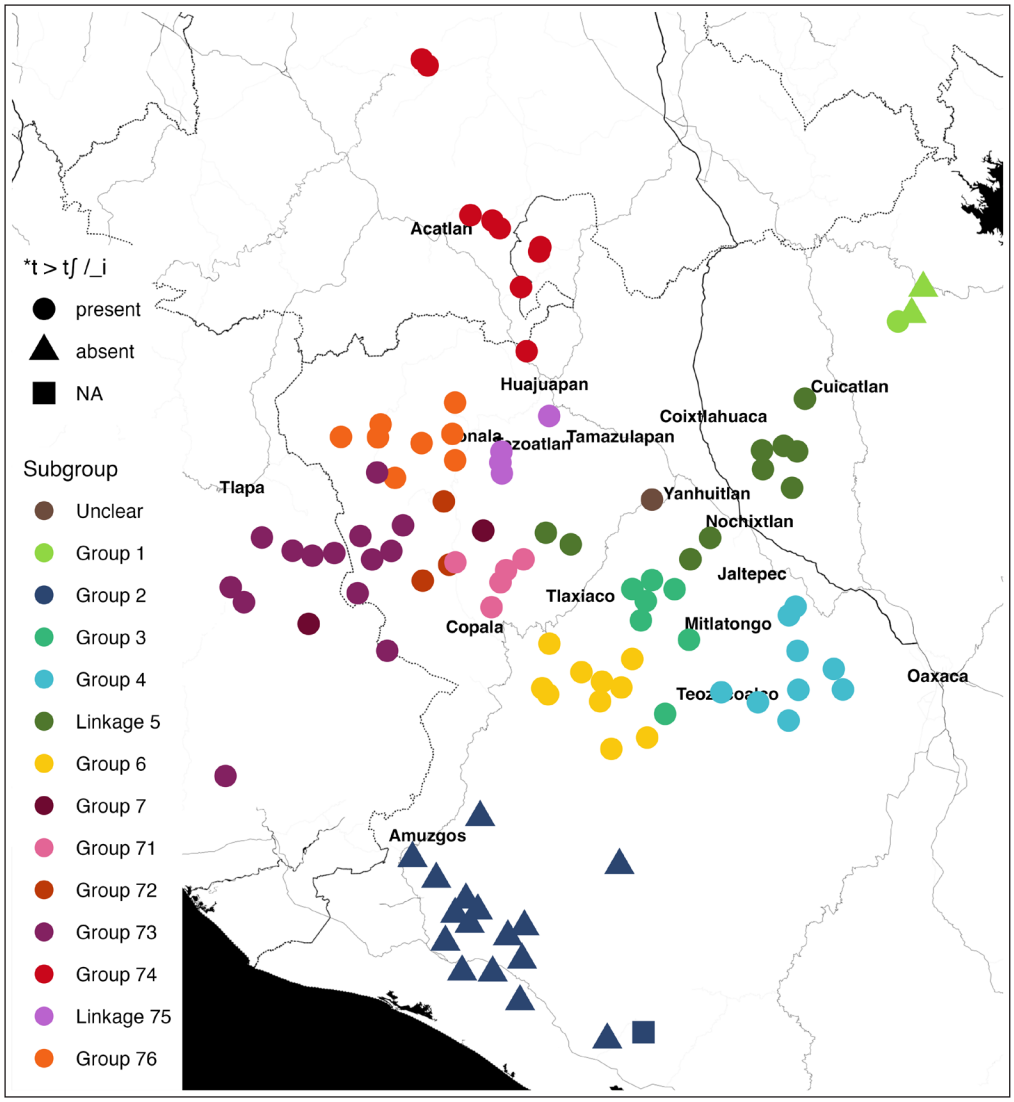

Figure 4 Presence (circle) and absence (triangle) of palatalization of *t before i, with subgroups (according to Auderset et al. 2023a) represented by colors.

7 CONCLUSION

The Mixtec sound change database provides a valuable and flexible resource for exploring the history of Mixtec languages and the relationships among them. It mobilizes comparable lexical data from 105 Mixtec doculects, by standardizing the representation of the data to IPA. Sound changes are identified in a bottom-up fashion by comparing the modern reflexes to proto-Mixtec reconstructions, which are reviewed and (minimally) revised as needed. This approach is empirical rather than theoretically driven. Components of the databases can all be linked to each other facilitating a range of possible uses of the data and the database. Importantly, this fine-grained approach is necessary due to the great number of Mixtec varieties, their overlapping and 'checkered' phonological histories, and the number of sound changes that have occurred across the language group. The design of the database allows it to be expanded in the future to address some of its current limitations. One such limitation concerns the restriction to segmental changes. Mixtec languages present many examples of suprasegmental change, a still poorly understood area in historical linguistics (Janda and Joseph 2003, 173; Campbell 2021a, 2021b), displaying different diachronic trajectories with respect to vowel nasality, a wide range of structural and typological diversity in their tone systems, and diachronic interaction between tone and other laryngeal features (Dürr, 1987; Pankratz & Pike, 1967). Another limitation concerns the relative chronology of the changes. Establishing a complete chronology of the sound changes identified in this study is important because it can help shed light on the complex migration history of Mixtec people. However, because of this complex history, this requires more in-depth research and detailed studies concerning specific changes, subgroups, and areas. While tone change and chronology are not included in the current version of the Mixtec sound change database, the methods used here provide a framework for doing so in the future. By applying methodology from autotypology

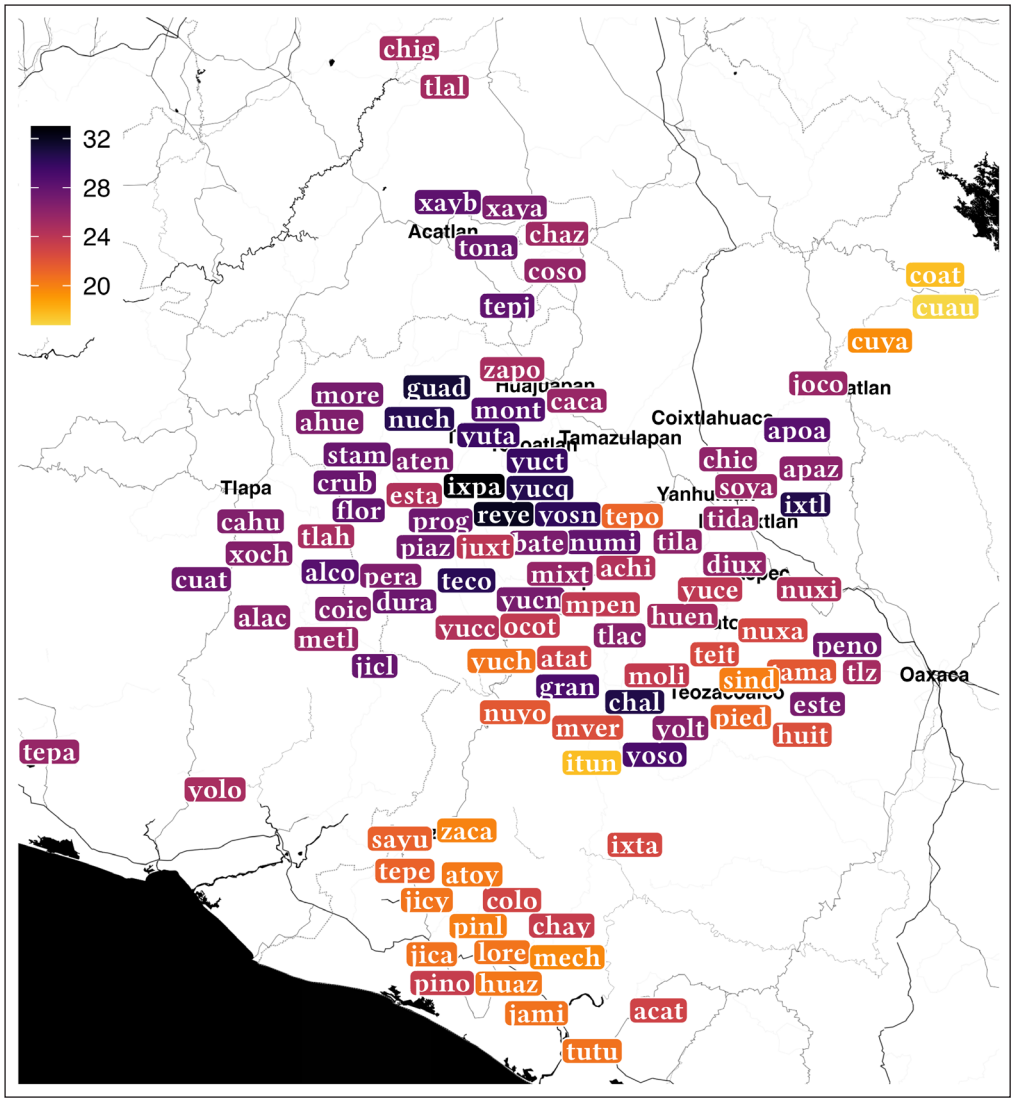

Figure 5 Percentage (number of changes present/total number of changes that are not NA in each variety) of changes across the sampled languages.

and providing the data collected and analyzed, we hope to encourage more detailed studies of subgroups within Mixtec and Mixtecan, but also studies on other language families.

ACKNOWLEDGEMENTS

We sincerely thank language workers who shared primary unpublished data for inclusion in the database (in alphabetical order by first name): Carmen Hernández Martínez, Griselda Reyes Basurto, Iní G. Mendoza, Jeremías Salazar, JN Martin, Jonathan D. Amith, Juvenal Solano, and Yésica Ramirez. The rest of the data are from published sources, and discussion of and references to these are included in the repository (in the metadata and orthography_to_ipa files).

FUNDING INFORMATION

This work was funded in part by the National Science Foundation award 1660355 to the University of California, Santa Barbara, PI: Eric W. Campbell and Mary Bucholtz; by the University of California, Santa Barbara Academic Senate Faculty Research Grant (2018–19), PI: Eric W. Campbell; and by the Endangered Languages Documentation Programme Small Grant 0566 (2019–2022) to the University of California, Santa Barbara, PI: Sandra Auderset.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

SA: Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing; EWC: Conceptualization, Supervision, Writing – original draft, Writing – review & editing.

AUTHOR AFFILIATIONS

Sandra Auderset orcid.org/0000-0002-4673-4814
Department of Linguistics, University of Bern, CH
Eric W. Campbell orcid.org/0000-0003-4100-5150
Department of Linguistics, University of California Santa Barbara, Santa Barbara, USA

REFERENCES

Auderset, S., Greenhill, S. J., DiCanio, C. T., & **Campbell, E. W.** (2023a). Subgrouping in a 'dialect continuum': A Bayesian phylogenetic analysis of the Mixtecan language family. *Journal of Language Evolution*, *8*(1). DOI: https://doi.org/10.1093/jole/lzad004

Auderset, S., Greenhill, S. J., DiCanio, C. T., & **Campbell, E. W.** (2023b, June). Supplementary Materials to "Subgrouping in a 'dialect continuum': A Bayesian phylogenetic analysis of the Mixtecan language family" (1.2 ed.). *Zenodo*. DOI: https://doi.org/10.1093/jole/lzad004

Bakker, D. (2010). Language sampling. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (pp. 100–127). Oxford University Press. DOI: https://doi.org/10.1093/oxfordhb/9780199281251.013.0007

Bickel, B. (2007). Typology in the 21 st century: major current developments. *Linguistic Typology*, *11*, 239–251. DOI: https://doi.org/10.1515/LINGTY.2007.018

Bickel, B. (2010). Capturing particulars and universals in clause linkage: a multivariate analysis. In I. Bril (Ed.), *Clause-hierarchy and clause-linking: the syntax and pragmatics interface* (pp. 51–101). Amsterdam: Benjamins. DOI: https://doi.org/10.1075/slcs.121.03bic

Bickel, B., Austin, P. K., Bond, O., Nathan, D., & **Marten, L.** (2011). Multivariate typology and field linguistics: a case study on detransitivization in Kiranti (Sino-Tibetan). In *Proceedings of the conference on language documentation and linguistic theory, 3*, 3–13. London: SOAS.

Bickel, B., & **Nichols, J.** (2002). Autotypologizing databases and their use in fieldwork. In *Proceedings of the international LREC workshop on resources and tools in field linguistics, Las Palmas* (Vol. 2627).

Bowern, C. (2018). Computational phylogenetics. *Annual Review of Linguistics*, 4, 281–296. DOI: https://doi.org/10.1146/annurev-linguistics-011516-034142

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

Campbell, E. W. (2017a). Otomanguean historical linguistics: Exploring the subgroups. *Language and Linguistics Compass*, *11*(7). DOI: https://doi.org/10.1111/lnc3.12244

Campbell, E. W. (2017b). Otomanguean historical linguistics: Past, present, and prospects for the future. *Language and Linguistics Compass*, *11*(4), 1–22. DOI: https://doi.org/10.1111/lnc3.12240

Campbell, E. W. (2021a). On Zapotecan glottal stop, and where (not) to reconstruct it. In M. Babel & M. A. Sicoli (Eds.), *Contact, structure, and change: A Festschrift in honor of Sarah G. Thomason* (pp. 349–382). Ann Arbor: Michigan Publishing.

Campbell, E. W. (2021b). Why is tone change still poorly understood, and how might documentation of less studied tone languages help? In P. Epps, D. Law, & N. Pat-El (Eds.), *Historical linguistics and endangered languages* (pp. 15–40). Routledge. DOI: https://doi.org/10.4324/9780429030390-3

Carroll, L. S. (2015). *Ixpantepec Nieves Mixtec word prosody* (Unpublished doctoral dissertation). University of California San Diego.

Castillo García, R. (2007). *Descripción fonológica segmental y tonal del mixteco de Yoloxóchitl, Guerrero* (Unpublished master's thesis). CIESAS, México, D.F.

Cruz, E., & **Woodbury, A. C.** (2014). Finding a way into a family of tone languages: The story and methods of the Chatino Language Documentation Project. *Language documentation & conservation*, *8*, 490–524.

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

Gerfen, H. J. (1996). *Topics in the phonology and phonetics of Coatzospan Mixtec* (Unpublished doctoral dissertation). University of Arizona.

Greenhill, S. J., Heggarty, P., & **Gray, R. D.** (2020). Bayesian phylolinguistics. In R. D. Janda, B. D. Joseph, & B. S. Vance (Eds.), *The Handbook of Historical Linguistics, 2*, 226–253. Wiley Blackwell. DOI: https://doi.org/10.1002/9781118732168.ch11

Hinton, L., Buckley, G., Kramer, M., & **Meacham, M.** (1991). Preliminary analysis of Chalcatongo Mixtec tone. In J. E. Redden (Ed.), *Papers from the American Indian Languages Conference, University of California,* Santa Cruz, July and August 1991 (pp. 147–155). Carbondale: Southern Illinois University.

Janda, R. D., & **Joseph, B. D.** (2003). On language, change, and language change–or, of history, linguistics, and historical linguistics. In B. D. Joseph & R. D. Janda (Eds.), *The Handbook of Historical Linguistics* (pp. 3–180). Oxford: Blackwell. DOI: https://doi.org/10.1111/b.9781405127479.2004.00002.x

Josserand, J. K. (1983). *Mixtec dialect history* (Unpublished doctoral dissertation). Tulane University.

Julián Caballero, J. (1999). La Academia de la Lengua Mixteca: espacios de reflexión compartida. *Cuadernos del Sur*, *14*, 129–139.

Kaufman, T. (2006). Oto-manguean languages. In K. Brown (Ed.), *Encyclopedia of language & linguistics* (Second Edition, pp. 118–124). Oxford: Elsevier. DOI: https://doi.org/10.1016/B0-08-044854-2/02286-0

Kaufman, T. (in press). Comparative Oto-Mangean grammar research: Phonology, aspect-mood marking, valency changers, nominalizers on verbs, numerals, pronouns, deictics, interrogatives, adpositionoids, noun classifiers, noun inflexion, compounds, word order, and diversification model. In S. Wichmann (Ed.), *Languages and linguistics of Mexico and Northern Central America: a comprehensive guide*. De Gruyter Mouton.

List, J.-M., Forkel, R., Greenhill, S. J., Rzymski, C., Englisch, J., & **Gray, R. D.** (2022). Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, *9*(1), 316. DOI: https://doi.org/10.1038/s41597-022-01432-0

List, J.-M., Greenhill, S. J., & **Gray, R. D.** (2017). The potential of automatic word comparison for historical linguistics. *PLoS ONE*, *12*(1), e0170046. DOI: https://doi.org/10.1371/journal.pone.0170046

List, J.-M., Walworth, M., Greenhill, S. J., Tresoldi, T., & **Forkel, R.** (2018). Sequence comparison in computational historical linguistics. *Journal of Language Evolution*, *3*(2), 130–144. DOI: https://doi.org/10.1093/jole/lzy006

Longacre, R. E. (1957). *Proto-Mixtecan*. Bloomington: Indiana University.

Longacre, R. E. (1961). Swadesh's Macro-Mixtecan hypothesis. *International Journal of American Linguistics*, 27(1), 9–29. DOI: https://doi.org/10.1086/464599

Macaulay, M., & **Salmons, J. C.** (1995). The phonology of glottalization in Mixtec. *International Journal of American Linguistics*, *61*(1), 38–61. DOI: https://doi.org/10.1086/466244

Mak, C., & **Longacre, R.** (1960). Proto-Mixtec phonology. *International Journal of American Linguistics*, *26*(1), 23–40. DOI: https://doi.org/10.1086/464551

McKendry, I. (2013). *Tonal association, prominence and prosodic structure in South-Eastern Nochixtlán Mixtec* (Unpublished doctoral dissertation). University of Edinburgh, Edinburgh.

Mendoza Ruíz, J. (2016). *Fonología segmental y patrones tonales del Tu'un Savi de Alcozauca de Guerrero* (Unpublished master's thesis). CIESAS, México, D.F.

Moran, S., Grossman, E., & **Verkerk, A.** (2021). Investigating diachronic trends in phonological inventories using BDPROTO. *Language Resources and Evaluation, 55*, 79–103. DOI: https://doi.org/10.1007/s10579-019-09483-3

North, J., & **Shields, J.** (1977). Silacayoapan Mixtec phonology. In W. R. Merrifield (Ed.), *Studies in otomanguean phonology* (pp. 21–33). Summer Institute of Linguistics and the University of Texas at Arlington.

Pankratz, L., & **Pike, E. V.** (1967). Phonology and morphotonemics of Ayutla Mixtec. *International Journal of American Linguistics*, *33*(4), 287–299. DOI: https://doi.org/10.1086/464980

Pike, E. V., & **Cowan, J. H.** (1967). Huajuapan Mixtec phonology and morphophonemics. *Anthropological Linguistics, 9*(5), 1–15.

Rensch, C. R. (1976). *Comparative Otomanguean phonology*. Bloomington: Indiana University Press.

Swanton, M. (2021). Un acercamiento a la ortografía dominica del mixteco de Teposcolula: un enfoque comparativo. In M. Swanton (Ed.), *Filología mixteca: Estudios sobre textos virreinales* (pp. 1–76). México, DF: Instituto de Investigaciones Filológicas UNAM.

Swanton, M., & **Mendoza Ruíz, J.** (2021). Observaciones sobre la diacronía del tono en el Tu'un Savi (mixteco) de Alcozauca de Guerrero. In F. Arellanes & L. Guerrero (Eds.), *Estudios lingüísticos y filológicos en lenguas indígenas mexicanas: Celebración de los 30 años del Seminario de Lenguas Indígenas*. Ciudad de México: Universidad Nacional Autónoma de México.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... others. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*(1), 1–9. DOI: https://doi.org/10.1038/sdata.2016.18

Witzlack-Makarevich, A., Nichols, J., Hildebrandt, K., Zakharko, T., & **Bickel, B.** (2022). Managing AUTOTYP data: Design principles and implementation. In A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Eds.), *The Open Handbook of Linguistic Data Management*. Cambridge: MIT Press. DOI: https://doi.org/10.7551/mitpress/12200.003.0061

TO CITE THIS ARTICLE:
Auderset, S., & Campbell, E. W. (2024). A Mixtec Sound Change Database. *Journal of Open Humanities Data*, 10: 24, pp. 1–13. DOI: https://doi.org/10.5334/johd.184

Submitted: 16 November 2023
Accepted: 08 February 2024
Published: 13 March 2024

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.