Journal of Open Humanities Data

Mixtec Sound Change Database 2.0: Integrating Tone Change

DATA PAPER

SANDRA AUDERSET

ubiquity press

ABSTRACT

This paper introduces and describes the second edition of the Mixtec Sound Change Database (Auderset & Campbell, 2024), which now includes a module on tone change. Tone change is an under-researched topic in historical linguistics and is virtually absent from cross-linguistic databases. The Mixtec languages from southern Mexico provide an ideal starting point for studies in this area, as they are all tonal and tone can be reconstructed to the proto-language. This updated version of the database thus serves as a model for including tonal data in synchronic and diachronic studies of languages and language families where this feature is present. This will support efforts towards a better understanding of suprasegmental change in Mixtec and beyond.

CORRESPONDING AUTHOR:

Sandra Auderset

Department of Linguistics, University of Bern, Bern, Switzerland

sandra.auderset@unibe.ch

KEYWORDS:

Tone change; sound change; Mixtec; comparative method

TO CITE THIS ARTICLE:

Auderset, S. (2025). Mixtec Sound Change Database 2.0: Integrating Tone Change. *Journal of Open Humanities Data*, 11: 37, pp. 1–6. DOI: https://doi.org/10.5334/johd.322

1 OVERVIEW

REPOSITORY LOCATION

The database is hosted on GitHub (https://github.com/SAuderset/MixteCaSo) so that it can be improved and expanded in the future. Publication versions such as this one are archived on Zenodo (DOI: 10.5281/zenodo.10143870).

CONTEXT

This database was initially created for Auderset (2022). The first open version is described in Auderset and Campbell (2024). The current, second version was created as part of the research for Auderset (2024).

Tone change is still understudied in historical linguistics. This pertains especially to tonal processes apart from tonogenesis (the emergence of tonal contrasts where there previously were none). Much of what we currently know about diachronic tonal processes draws on studies of Southeast Asian languages, where tone emerged only relatively recently from segments (Dockum, 2019; Ferlus, 2004; Gedney, 1972; Joseph & Burling, 2001, among others). Mixtec languages are part of the Mixtecan branch of Otomanguean, a language family where tone is old and likely dates back to Proto-Otomanguean (Campbell, 2017; Rensch, 1976). This makes it an ideal choice for investigating tone change alongside segmental change. Furthermore, there is earlier work on tone reconstruction and tone diachrony (Dürr, 1987; Swanton & Mendoza Ruiz, 2021) in Mixtec. Recent years have also seen increased documentation efforts, such that more materials are now available for comparative tonal studies in this language family. This second version of the Mixtec Sound Change Database aims at making these materials available to a wider audience, as well as working towards closing the research gap concerning the diachrony of tone.

2 METHOD

STEPS

Comparing tones is generally challenging, even within a language family or subgroup. For diachronic studies such as reconstruction, it is important that we compare tones at the same level of abstraction – just as with segments. This means that the tones represented in the database are contrastive, i.e. they are tonemes rather than phonetic tones. Tone correspondences were established based on the comparative method, following the methodology for segmental correspondences and changes illustrated in Auderset and Campbell (2024). The correspondences, which form the basis for identifying and coding the tone changes, are established based on cognate sets. The tonal reconstructions by Dürr (1987) and Swanton and Mendoza Ruiz (2021) served as a starting point. A detailed explanation of the coding process is outside the scope of this paper, but more details and examples can be found in Auderset (2024, 11–18) and in the supplementary materials to the database on GitHub (at definitions/tone_standardization.pdf[1]). Because Proto-Mixtec was already tonal, it is often not possible to establish segmental origins or conditioning environments of correspondence sets. I thus refrained from positing specific tones or glottal stops of preceding constituents that would have conditioned some of the correspondence sets. Instead, I assign labels to the sets that stand in for environments that can later be elaborated upon. These labels are explained in more detail in Auderset (2024, 15–17). Such advances will depend on a better comparative understanding of floating tones and tone sandhi phenomena in Mixtec on a synchronic and diachronic level.

SAMPLING STRATEGY

Due to the still wide-spread practice of omitting tone from descriptions of tonal languages or representing it only in a select few cases, the sample for the tone module is restricted. I included all Mixtec varieties for which tonal contrasts were marked systematically and in a reliable way.

1 Path to file: https://github.com/SAuderset/MixteCaSo/blob/b38621d4ac309e23fde616b5424347586a8dbf2c/definitions/tone_standardization.pdf.

Out of the 105 varieties for which segmental changes can be annotated, only 46 have tones marked on more than a few entries. Three varieties (Cahuatache, Diuxi, and Abasolo del Valle Mixtec) could not be coded for tone changes due to difficulties in the interpretation of the tone values in the sources. The data for Cahuatuache Mixtec are from an older source (Schultze-Jena, 1938) and the tone notation is inconsistent, such that regular correspondences could not be established. For Diuxi Mixtec, several materials are available that use differing analyses of the tone system, ranging from two tonemes in combination with stress (Pike & Oram, 1976) to four tonemes and no stress (Daly, 1978). Further complicating the matter are various tone sandhi processes and the fact that Pike and Oram (1976) provide surface tones but Daly (1978) underlying tones. The data are reproduced in Kuiper and Oram (1991) and Dürr (1987) but without standardization. An in-depth analysis and reconciliation of these different notations and analyses is outside the scope of the current work and the variety thus had to be excluded. In the materials on Abasolo del Valle (Galindo Sánchez, 2009), tones are marked throughout but I could not establish regular correspondences with other varieties, even closely related ones. I thus chose not to include this variety in the study. The resulting 43 languages belong to six of the seven larger subgroups identified in Auderset et al. (2023) and to 10 out of the 12 dialect areas identified by Josserand (1983). They are illustrated in Figure 1. Even though this is a convenience sample that does not cover all subgroups, the varieties included provide good coverage of the attested diversity of tonal systems in Mixtec.

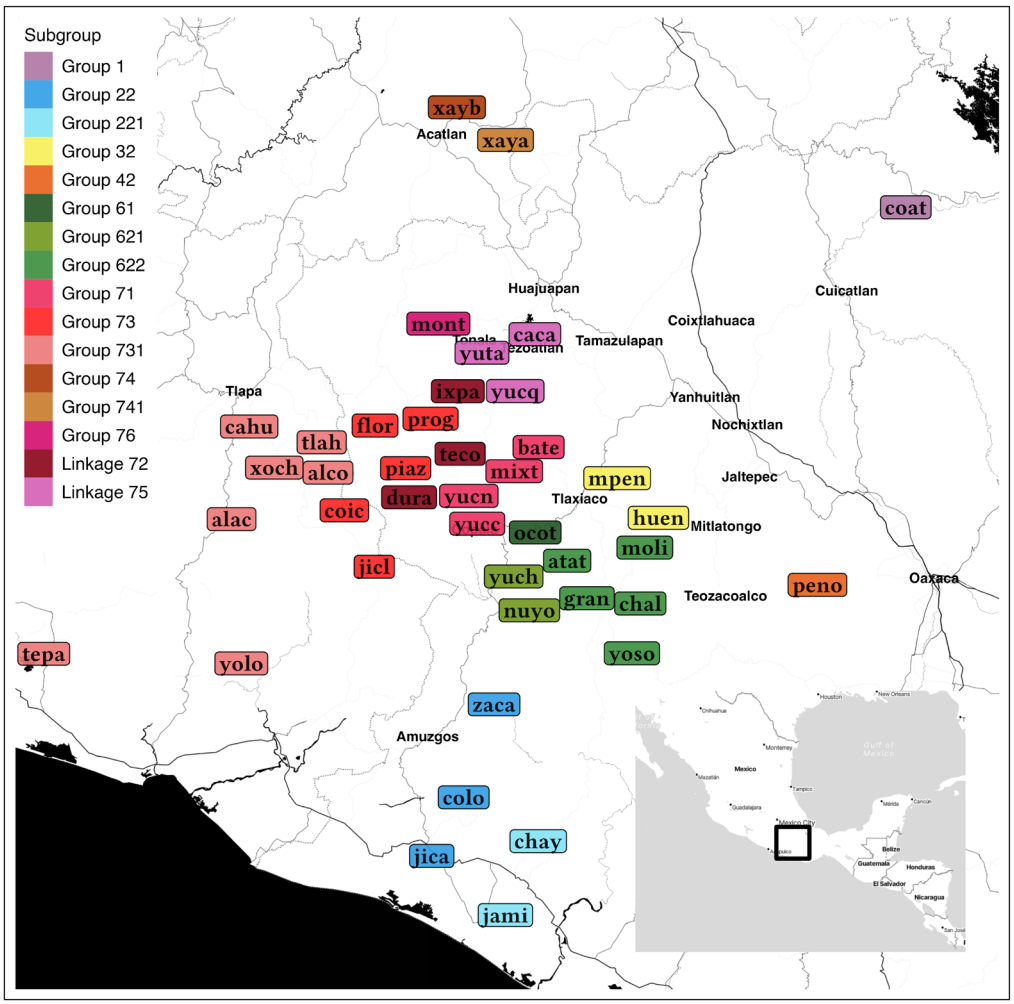

Figure 1 Map of Mixtec varieties for which tone changes could be coded with their subgroup affiliation (according to Auderset et al. 2023). The inset shows the location of the detailed map within Mesoamerica. Reproduced from Auderset 2024, which can be consulted for more details.

NOTATION

The IPA offers two principal ways of displaying tone: diacritics that are placed above the tone-bearing unit or tone bars, with the latter method suggested as the preferred one. Tone diacritics are useful in practical orthography but not well suited to alignment for comparative purposes or for computational processing. For such purposes, it is more useful to represent the tone after the tone-bearing unit as its own character (even if this does not reflect phonetic reality). Therefore, I use Chao's tone numbers (Chao, 1930), 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. The interval between the lowest and highest pitch is assumed to correspond roughly to an augmented fifth (Chao, 1930). Contour tones are represented as combinations of these levels. A high to low falling tone, for example, is noted as 51.

3 DATASET DESCRIPTION

REPOSITORY NAME

MixteCaSo 2.0

OBJECT NAME

SAuderset/MixteCaSo-2.0.0.zip

FORMAT NAMES AND VERSIONS

tsv, PDF, R, Rmd

CREATION DATES

2022-01-01 to 2024-02-01

DATASET CREATORS

Sandra Auderset (creator, annotater), Eric W. Campbell (advisor)

LANGUAGE

English, Spanish

LICENSE

CC-BY-SA-4.0

PUBLICATION DATE

2025-03-06

4 REUSE POTENTIAL

The data presented in the tone module serves various audiences and can be used to explore research questions in multiple subfields of linguistics. Comparative data on tone that is standardized and covers more than a handful varieties of a language family is difficult to come by and as such, the database also serves as a model for future resources on other language families. One obvious avenue of research is in historical linguistics and diachronic linguistics more generally. Tone change is still an understudied area not least because comparative data is often lacking (Campbell, 2021). This database allows for an updated reconstruction of Mixtec tone (cf. Dürr 1987; Swanton and Mendoza Ruiz 2021), but more importantly for a comparison with segmental change which is coded in the same way. Auderset (2024) explores the rates of change and the potential for subgrouping in segmental and tonal changes. The database also provides detailed information for microtypological studies on tone. Tone has also been a neglected topic in (phonological) typology (Moran et al., 2023) and regional or family-specific studies show that the internal diversity of tonal phenomena is often underrepresented in cross-linguistic surveys (Brunelle & Kirby, 2016). One limitation of the database and the tone module specifically is that it relies on a mix of published data and data that has been collected recently by the author or colleagues through recordings of naturalistic speech. This means that the confidence level with respect to the tone analysis is not the same across all sources. However, this limitation is not specific to this data set, but inherent in most cross-linguistic databases.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank all the Ñuu Savi and linguists collaborating with them for sharing their data to make this database possible. Special thanks go to Carmen Hernández Martínez, Griselda Reyes Basurto, Inî G. Mendoza, Jeremías Salazar, JN Martin, Jonathan D. Amith, Juvenal Solano, and Yésica Ramirez for contributing primary unpublished data for inclusion in the database. The rest of the data are from published sources, and discussion of and references to these are included in the supplementary material on GitHub.

FUNDING STATEMENT

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

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR ROLES

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

AUTHOR AFFILIATIONS

Sandra Auderset orcid.org/0000-0002-4673-4814
Department of Linguistics, University of Bern, Bern, Switzerland

REFERENCES

Auderset, S. (2022). *Confronting challenges in historical linguistics: Quantitative approaches to dialect area subgrouping and tone change in Mixtec* (Doctoral dissertation, University of California Santa Barbara). Retrieved from https://www.proquest.com/dissertations-theses/confronting-challenges-historical-linguistics/docview/2728523826/se-2

Auderset, S. (2024). Rates of change and phylogenetic signal in Mixtec tone. *Language Dynamics and Change*, *14*(1). https://doi.org/10.1163/22105832-bja10031

Auderset, S., & **Campbell, E. W.** (2024). A Mixtec sound change database. *Journal of Open Humanities Data*, *10*(1). https://doi.org/10.5334/johd.184

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

Brunelle, M., & **Kirby, J.** (2016). Tone and phonation in Southeast Asian languages. *Language and Linguistics Compass*, *10*(4), 191–207. https://doi.org/10.1111/lnc3.12182

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

Campbell, E. W. (2021). 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). New York: Routledge. https://doi.org/10.4324/9780429030390-3

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

Daly, J. P. (1978). Notes on Diuxi Mixtec tone. *Work Papers of the Summer Institute of Linguistics, University of North Dakota Session*, *22*(1). https://doi.org/10.31356/silwp.vol22.03

Dockum, R. (2019). *The tonal comparative method: Tai tone in historical perspective* (Unpublished doctoral dissertation). Yale University.

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

Ferlus, M. (2004). The origin of tones in Viet-Muong. In S. Buruspat (Ed.), *Papers from the eleventh annual meeting of the Southeast Asian Linguistics Society 2001* (pp. 297–313). Tempe, Arizona: Arizona State University Programme for Southeast Asian Studies Monograph Series Press.

Galindo Sánchez, B. (2009). *Vocabulario Básico Tu'un Savi – Castellano* (1st ed.). Xalapa, Veracruz: Academia Veracruzana de las Lenguas Indígenas.

Gedney, W. J. (1972). A checklist for determining tones in Tai dialects. In M. E. Smith (Ed.), *Studies in linguistics in honor of George L. Trager* (pp. 423–437). The Hague: Mouton.

Joseph, U. V., & **Burling, R.** (2001). Tone correspondences among the Boro languages. *Linguistics of the Tibeto-Burman Area*, *24*(2), 41–55. https://doi.org/10.32655/LTBA.24.2.02

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

Kuiper, A., & **Oram, J.** (1991). A syntactic sketch of Diuxi-Tilantongo Mixtec. In C. H. Bradley & B. E. Hollenbach (Eds.), *Studies in the syntax of Mixtecan languages* (Vol. 3, pp. 185–408). Summer Institute of Linguistics and the University of Texas at Arlington.

Moran, S., Easterday, S., & **Grossman, E.** (2023). Current research in phonological typology. *Linguistic Typology*, *27*(2), 223–243. https://doi.org/10.1515/lingty-2022-0069

Pike, E. V., & **Oram, J.** (1976). Stress and tone in the phonology of Diuxi Mixtec. *Phonetica*, *33*(5), 321–333. https://doi.org/10.1159/000259780

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

Schultze-Jena, L. (1938). *Bei den Azteken, Mixteken und Tlapaneken der Sierra Madre del Sur von Mexico* (Vol. 3). Jena: Gustav Fischer.

Swanton, M., & **Mendoza Ruiz, 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.

TO CITE THIS ARTICLE:
Auderset, S. (2025). Mixtec Sound Change Database 2.0: Integrating Tone Change. *Journal of Open Humanities Data*, 11: 37, pp. 1–6. DOI: https://doi.org/10.5334/johd.322

Submitted: 07 March 2025
Accepted: 19 April 2025
Published: 11 June 2025

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