Title

Confronting challenges in historical linguistics: Quantitative approaches to dialect area subgrouping and tone change in Mixtec

Author

Auderset, Sandra

Publication Date

2022

University of California
Santa Barbara

Confronting challenges in historical linguistics: Quantitative approaches to dialect area subgrouping and tone change in Mixtec

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Linguistics

by

Sandra Auderset

Committee in charge:

Professor Eric W. Campbell, Chair
Professor Marianne Mithun
Professor Lynn Hou
Professor Simon J. Greenhill, University of Auckland & Max Planck Institute for Evolutionary Anthropology, Leipzig
Professor Russell D. Gray, Max Planck Institute for Evolutionary Anthropology, Leipzig

September 2022

The Dissertation of Sandra Auderset is approved.

Professor Marianne Mithun

Professor Lynn Hou

Professor Simon J. Greenhill, University of Auckland & Max Planck Institute for Evolutionary Anthropology, Leipzig

Professor Russell D. Gray, Max Planck Institute for Evolutionary Anthropology, Leipzig

Professor Eric W. Campbell, Committee Chair

September 2022

Confronting challenges in historical linguistics: Quantitative approaches to dialect area subgrouping and tone change in Mixtec

by

Sandra Auderset

To the Ñuu Savi
past, present, and future

Acknowledgements

First and foremost, I thank all the Ñuu Savi who I met over the past years in California and in Oaxaca. Without their enthusiasm, patience, and support this dissertation would not have been possible. My sincere thanks to Carmen Hernández Martínez, the best collaborator one could ever wish for. It's not an overstatement to say that without her I probably would not have written my dissertation on Mixtec in the first place. Special thanks also to Carmen's mother, Catalina Martínez Ramirez, who hosted us during our first stay in Oaxaca and never turned down an opportunity for a recording or teaching moment, be it on Mixtec or any other life skill. I'm also grateful to all the people I met at MICOP (Mixteco/Indígena Community Organizing Project, Ventura County, California) during my participation in the community language classes 2017-2019. Thank you to Griselda Reyes Basurto, Yésica Ramirez, and Juvenal Solano for sharing your language and so much more. I learned a lot during my two stays in San Martín Duraznos about Ñuu Savi culture and language and I extend my gratitude to everyone I have met there, especially Flor Melina Herón Reyes, Reina Martínez Rendón, and Pedro Pérez Mendoza. Many thanks to my UCSB Mixtec colleagues for all the valuable insights and discussions, especially to Inî G. Mendoza and Simon L. Peters.

I am grateful to all the community members and linguists who shared their data with me and allowed me to use and publish it (some already mentioned above): Braulio Becerra Roldán (Santo Domingo Huendio Mixtec), Carmen Hernández Martinez (San Martin Duraznos Mixtec), Christian DiCanio (Itunyoso Triqui), Fidel Hernández Mendoza (Chicahuaxtla Triqui), Griselda Reyes Basurto (Tlahuapa Mixtec), Inî G. Mendoza and Simon L. Peters (Piedra Azul Mixtec), Jeremías Salazar (Yucunani Mixtec), JN Martín (El Jicaral Mixtec), Juvenal Solano (San Sebastian del Monte Mixtec), Yésica Ramirez (La Batea Mixtec), Rey Castillo Garcia and Jonathan Amith (Yoloxochitl Mixtec).

At UCSB, I was lucky to be surrounded by wonderful colleagues. Thank you to Adrienne and Jamaal, the best cohort buddies ever - I learned so much from you. Many thanks also to my Zoom writing group, who kept me sane during the pandemic and the last months of dissertation writing: Adrienne, Alexia, Giorgia, Jamaal, and Karen. I don't think I would have made it without your support.

Thanks to my committee members Marianne, Lina, Simon and Russell, for the support and encouragement and especially for working with me to make a very crunched timeline possible. I'm grateful to committee chair Eric Campbell for introducing me to the wonderful world of Mixtec linguistics and language documentation. I would like to thank Russell Gray for funding me for the last three years of my program through the MPI. My stay at the MPI has greatly benefited my research and I'm particularly grateful to Simon Greenhill for working with me through all my phylogenetics and other methodological questions. Thanks also to my MPI colleagues for the good times and special shout-out to Hedvig for the R support.

I extend my gratitude to my parents, Romy Auderset and David Steinore, who supported me throughout this adventure no matter where it took me and always believed in me. Finally, thanks to Adam for listening to all my complaints and for everything else.

Curriculum Vitæ
Sandra Auderset

Education

2022	Ph.D. in Linguistics, University of California, Santa Barbara
2015	M.A. in General Linguistics, University of Zurich
2013	B.A. in Comparative Indo-European Linguistics, University of Zurich

Professional Employment

2019-2022: Doctoral researcher, Max Planck Institute for Evolutionary Anthropology, Leipzig

2018: Instructor of Record, LING 104: Statistical Methodology, Department of Linguistics, University of California Santa Barbara

2016-2019: Teaching Assistant, various classes, Department of Linguistics, University of California Santa Barbara

2015-2016: Research Assistant, University of Zurich

2011-2015: Student Assistant, University of Zurich

Publications

Auderset, Sandra, Carmen Hernández Martínez & Albert Ventayol-Boada. accepted. Constituency in Tù'un Ntá'ví (Mixtec) of San Martín Duraznos. in: Tallman, Auderset, Uchihara (eds.) *Constituency in the Americas.* Language Science Press

Tallman, Adam J.R. & Sandra Auderset. 2022. Measuring and assessing indeterminacy and variation in the morphology-syntax distinction. *Linguistic Typology* published online March 2022

Auderset, Sandra. 2021. The antipassive and its relationship to person markers. In: Janic, Katarzyna & Alena Witzlack-Makarevich (eds.). *Antipassive: Typology, diachrony, and related constructions.* Typological Studies in Language 130. Amsterdam: Benjamins

Auderset, Sandra. 2020. Interrogatives as relativization markers in Indo-European. Diachronica 37(4)

Lester, Nicholas A., Sandra Auderset & Phillip G. Rogers. 2018. Case inflection and the functional indeterminacy of nouns: A cross-linguistic analysis. *Proceedings of the 40th Annual Conference of the Cognitive Science Society.* 2029–2034

Widmer, Manuel, Sandra Auderset, Johanna Nichols, Paul Widmer & Balthasar Bickel. 2017. NP recursion over time: evidence from Indo-European. *Language* 93(4), 799–826

Abstract

Confronting challenges in historical linguistics: Quantitative approaches to dialect area subgrouping and tone change in Mixtec

by

Sandra Auderset

This dissertation investigates the dynamics of language change in the Mixtec languages, a branch of Mixtecan (Otomanguean). Spread over a large mountainous area in southern Mexico, the history of this language family is characterized by repeated migrations, diversified local varieties, and ongoing inter-varietal contact, leading to its characterization as a dialect continuum. Mixtec languages are also well known for their intricate systems of grammatical and lexical tone, which have to be reconstructed to Proto-Mixtec and beyond.

The first study applies a Bayesian phylogenetic model to lexical cognacy data of 137 Mixtecan varieties. This model recovers the three main branches of Mixtecan (Mixtec, Triqui, and Cuicatec) as well as a number of well-supported higher- and lower-level groups within Mixtec. It also identifies varieties that are only loosely connected to the rest of the family and thus reveals that there are both wave-like and tree-like diversification processes at work and shows that subgrouping is possible and informative in a language family characterized as a dialect continuum.

The second study provides a comprehensive account of sound changes in Mixtec. The sound changes are coded as fine-grained variables in a framework adapted from multivariate typology, and analyzed using computational methods. This allows for high accuracy and detail with regards to the sound changes but also for making visible larger patterns and tendencies. The latter show that the distribution of Mixtec sound changes generally aligns well with the subgroups proposed in the first study.

The third study addresses the question of whether tones can be used for subgrouping and whether or not they change faster than segments. Building on the same typological framework as for the second study, I show that tone change can be analyzed in the same way as segmental change. I apply quantitative methods to calculate phylogenetic signal and rates of tonal versus segmental change based on the family tree obtained in the first study. The results show that tone change does not differ from segmental change in any meaningful way; that is, tones also show phylogenetic signal and their rates of gain and loss are not different from those of segments.

Each study is designed to address the research questions in a empirical way, by including all sufficiently reliable and available Mixtec language data, which will be shared in public repositories along with all aspects of data handling, coding, and scripts used in the analyses.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This dissertation brings together three studies on historical aspects of Mixtec languages. Each study focuses on an important topic within Mixtec linguistics that touches on larger issues within the field of historical linguistics and language change more broadly. I combine rigorous qualitative work with innovative quantitative approaches. Particular attention is paid to creating re-usable databases and providing replicable scripts for the analyses and visualizations. The basis of all three studies is a large database of over 1000 annotated cognate sets across 137 Mixtecan languages from published sources but also from ongoing documentation projects. Each study is designed to address the research question in an empirical way, by making the data and scripts available in public repositories, including detailed supplementary materials on the data gathering and analysis process.

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 languages are most closely related to the Cuicatec and Triqui languages, and these three groups together comprise the Mixtecan language family (Longacre 1957, 1961), which in turn is considered one of the major branches of the large Otomanguean family of Mesoamerica (Rensch 1976, Kaufman 2006, Campbell

2017b). The highly diversified Mixtec family is often characterized as a dialect continuum consisting of chains of mutually intelligible varieties (Jiménez Moreno 1962, Josserand 1983:457-458). This has led scholars to doubt the applicability of the comparative method and subgrouping with a tree model in this language family: Bradley & Josserand (1982:303) believed that dialect areas like those characterizing the Mixtec branch are not subject to the same processes of language change as 'traditional' language families and that we thus cannot go beyond positing broad dialect areas based on overlapping and competing isoglosses. Reconstructions of Proto-Mixtecan often included only a small number of Mixtec varieties and thus provided an incomplete picture of Proto-Mixtec (Longacre 1957, Gudschinsky 1959, Arana Osnaya 1960, Longacre 1962, Swadesh 1967). However, later work drawing on a large-scale language survey carried out in the Mixteca in the late seventies (Josserand 1983:130) added considerable detail to the reconstruction of proto-Mixtec (Bradley & Josserand 1982) and to the diversification scenario of this language family based on vowel isoglosses (Josserand 1983). Mixtec languages are all tonal and tone carries a high functional load both in lexicon and grammar. Tones in Mixtec are old and must be reconstructed to proto-Mixtec and proto-Mixtecan and most probably as far back as proto-Otomanguean (Campbell 2017b, Rensch 1976). Here, too, we can draw on earlier research on Mixtec tone reconstruction and tone change (Dürr 1987, Swanton & Mendoza Ruíz 2021). Recent years have seen an increase in the documentation of Mixtec languages, both quantitatively and qualitatively.[1] There is thus much more data available than what existed at the time of much of the earlier historical work. This increase in data has not been met with a similar increase in large-scale comparative Mixtec work and this thesis aims at providing a starting point for filling this gap.

Chapter 1 deals with subgrouping in a dialect continuum. The question of whether processes of diversification are fundamentally different in dialect continua from those in other

[1] Of the 358 works listed in Glottolog (Hammarström et al. 2022) for Mixtec about two thirds were published after the year 2000.

languages is part of a long-standing debate in the field of historical linguistics (see Heggarty et al. 2010, François 2015, Kalyan & François 2018, Jacques & List 2019 for recent discussion, among others). In the standard tree model, each split based on shared innovations is assumed to completely separate the daughter languages. In a dialect continuum, this assumption does not hold because the speech communities remain in contact with each other and shared innovations only permeate a part of the new subgroup (Bloomfield 1933:317). This problematizes the application of the comparative method, which rests on the principle of subgrouping by exclusive, shared innovations (Rankin 2003). One approach to deal with this issue has been to abandon tree models in such cases and instead advocate alternatives like the wave model (Schuchardt 1900, Schmidt 1872), network models (Heggarty et al. 2010, Willems et al. 2016), or historical glottometry (Kalyan & François 2018). Advocates of such alternative models argue that they better represent the developments within dialect continua (Ross 1996). Other scholars maintain that the tree model, if applied and interpreted correctly, is still appropriate in such circumstances (Jacques & List 2019), because ongoing contact between related languages does not preclude an investigation into the languages' genealogical relationships (Bowern 2013:426-427). Bayesian phylogenetics methods provide such a quantification and nuanced view. Originally developed for modeling processes of biological evolution, these methods have gained traction in linguistics in the past decade. Bayesian phylogenetic models make use of Bayes' Theorem to infer a sample of trees that have a high probability of explaining the distribution of the data under a given model of change (for a more technical explanation see Greenhill et al. 2020:228-232). The advantage of such models over constructing trees by hand is that they are more explicit and instead of providing a single tree, they result in a posterior distribution of trees, which quantify uncertainty in groupings. This latter point is especially important in cases where language families are characterized as dialect continua, since this uncertainty can show us where and when we find more or less tree-like processes. I apply a Bayesian phylogenetic model using the annotated cognate sets to explore the internal structure of the Mixtecan language family. This model recovers the three main branches of Mixtecan (Mixtec, Triqui, and Cuicatec) as well as a number of well-supported higher- and lower-level groups within Mixtec. It also identifies varieties that are only loosely connected to the rest of the family and thus reveals that there are both wave-like and tree-like processes at work and shows that subgrouping is possible and informative in a language family characterized as a dialect continuum.

Chapter 2 provides a comprehensive, updated overview of the segmental sound changes found in Mixtec and their distribution across the languages of this family. The focus on lies on evaluating earlier proposals of reconstructions and sound changes and assessing to what extent the distribution of sound changes aligns with previously proposed dialect areas (Josserand 1983) and subgroups (Auderset et al. submitted). This is challenging because of the large number of languages involved, which undergo many similar, but not identical changes. This has led to a situation in which the phoneme inventories of contemporary varieties look very similar, but the changes that led to this outcome are not necessarily the same (Josserand 1983:459). Given the characterization of Mixtec as a dialect continuum, we generally expect sound changes to show many local but overlapping distributions of the kind that are often found represented in isogloss maps. To investigate this assumption, we cannot cherry-pick changes deemed important, because there is no *a priori* way of knowing which and how many changes proceed in this way. To address this challenge, I combine methodology from historical linguistics and typology, creating reproducible and re-usable interlinked databases that allow for tracking the intricate patterns of sound change in this language family. I code all regular sound changes found in the data set such that they account for all the contemporary data. To explore general tendencies in the distribution of the changes, I draw on quantitative data analysis and visualization techniques. I calculate an association measure that evaluates how well the distribution of each sound change aligns with subgroups proposed with cognacy data. I use hierarchical clustering and a principal components analysis to capture the relationships between the languages based on the sound changes. All three methods point in the same direction, namely that sound changes generally align well with the subgroups arrived at in Chapter1. The relationships based on the sound changes recover many groupings similar to the cognacy-based subgroups.

Chapter 3 investigates the question of whether tones change faster than segments. Suprasegmental change in general and tone change specifically is still understudied in historical linguistics (Campbell 2021), despite the advances in data and methods of the field in other areas. The question of whether tone change proceeds in the same or a similar way as segmental change is important because the majority of the world's languages are tonal (Yip 2002). Observations that tone can vary quite drastically even among closely related varieties (Cahill 2011, Beam de Azcona 2007, Morey 2005, Dürr 1987) has led to the assumption that tones are inherently unstable and thus difficult to reconstruct, but we lack empirical studies on this question. Complicating matters is a similar gap in tone typology, that is in the systematic comparison of tone systems beyond simple classifications by the number of tones. I address this issue by using methods from multivariate typology, as with segmental changes. This entails the creation of interlinked databases of tonal and segmental sound changes. This fine-grained sound change data serves as the input for the calculation of phylogenetic signal with the metric *D* and estimated rates of gain and loss with a Hidden Markov Model across a posterior sample of trees from the first study. The results show that the majority of tone changes show phylogenetic signal and that they do not change at a faster rate than segments.

Together, these three chapters provide some answers to and significant advances in our understanding of the history of Mixtec(an) languages with regards to subgrouping, sound change, and tonal diachrony. Moreover, the questions of subgrouping within dialect continua and the nature of tone change are some of the major areas in need of advancement in historical linguistics more generally. Finally, the history of Mixtec languages is of great importance to the communities to whom these language belong and whose history and culture they encode.

In the face of pressures towards language shift that many Mixtec communities face (as other indigenous communities worldwide), better understanding of Mixtec language relations has potential for improving the design and application of materials for language maintenance and pedagogy.

Chapter 2

Subgrouping in a ‘dialect continuum’: A Bayesian phylogenetic analysis of the Mixtecan language family

Disclaimer

A slightly modified version of this chapter was submitted to the Journal of Language Evolution as: Auderset, Sandra, Simon J. Greenhill, Christian T. DiCanio & Eric W. Campbell. Subgrouping in a ‘dialect continuum’: A Bayesian phylogenetic analysis of the Mixtecan language family.
Author contributions: SA: Conceptualization, Data curation, Formal analysis, Methodology, Writing - original draft, Writing - review and editing; SJG: Formal analysis, Methodology, Writing - review and editing; CTD: Data curation, Writing - review and editing; EWC: Data curation, Writing - original draft, Writing - review and editing.

2.1 Introduction

The comparative method of historical linguistics is still seen as the most reliable method for uncovering genealogical language relationships. It works well in many cases when languages are expanding and diversifying, but is less successful in recovering historical relationships in dialect continua. In dialect continua, speech communities do not neatly separate from each other, but rather remain in on-going contact. This leads to differential diffusion between varieties, rendering a simple ‘tree-like’ model of their history inappropriate at best (Ross 1996) or invalid at worst (Kalyan & François 2018). It has thus sometimes been suggested that the comparative method of historical linguistics or Bayesian phylogenetic methods cannot be applied to dialect continua and dialect chains, given that the varieties remain in contact and there would be too much noise from internal loans (Ross 1996). This had led the critics of the simple tree model to the adoption of alternative approaches, such as the ‘wave model’ (proposed by Johannes Schmidt, cf. Anttila 1989), ‘network models’ (Willems et al. 2016), or ‘historical glottometry’ (Kalyan & François 2018). Defenders of the tree model point out that the impact and prevalence of contact on genealogical relationships is often overestimated and poses less of a problem for comparative reconstruction than is generally assumed . Apart from very specific circumstances, extensive language contact does not preclude an investigation into the languages’ genealogical relationships (Bowern 2013:426-427). However, a more fruitful way forward is to ask when and where language relationships are more like trees or more like waves, and to use approaches that can quantify and incorporate both patterns. Our task is thus not to define the complete history of a language family as pertaining to one or the other, but rather to identify which parts of its history can be described as tree-like and where we find conflicting signal that can be described as wave-like. Recently developed Bayesian phylogenetic methods provide such a way forward. Rather than providing a single tree estimate of the history of a language family, they provide a posterior probability distribution of many trees. Each tree in the posterior is a particular estimate of underlying history, and by aggregating over the trees we can identify which groupings are well supported by the data and which are noisier and show more conflict. These methods are also robust to even high levels of borrowing (Greenhill et al. 2009).

The Mixtecan language family of Mexico provides an excellent case study for the application of Bayesian phylogenetics for several reasons. First, there have been conflicting opinions about which languages belong to Mixtecan in the first place (see the debate between Swadesh 1960 and Longacre 1961, summarized in Section 2.2). Second, the highly diversified Mixtec subgroup within Mixtecan is often characterized as a dialect continuum consisting of chains of mutually intelligible varieties (Jiménez Moreno 1962, Josserand 1983:457-458). Third, documentation of Mixtec varieties has significantly increased in recent years, making it possible to re-evaluate earlier proposals and apply phylogenetic methods to a large and updated Mixtecan data set. Fourth, and finally, doubts about the possibility of successful subgrouping within Mixtec proper have been explicitly expressed in seminal work by Bradley & Josserand (1982:303). Like many other scholars, they believed that dialect areas like those characterizing the Mixtec branch are not subject to the same processes of language change as ‘traditional’ language families:

> “El problema es la aparente violación o infracción de las expectativas ideales del método comparativo en la linguística histórica (...). Pero este modelo presupone una efectiva separación de los grupos, después de un cambio que los divide. En contraste, en la Mixteca (...) vemos una situación más dinámica en el desarrollo de las familias lingüísticas.” (Bradley & Josserand 1982:303)[1]

We show that the application of Bayesian phylogenetics to Mixtecan produces valuable

[1] English version: The problem is the apparent violation or infringement of the expectations and ideals of the comparative method in historical linguistics (...). But this model assumes an actual separation of the groups, once a change has divided them. In contrast, in the Mixteca, we see a more dynamic situation concerning the development of the language families. (translation ours)

results and new insights with respect to subgrouping beyond what the comparative method and dialect geography can provide. The Mixtecan language family, classified by most as belonging to the Otomanguean stock of Mesoamerica (Rensch 1976, Campbell 1997, Kaufman 2006, Campbell 2017a), spreads over large parts of the southern Mexican state of Oaxaca and into Puebla and Guerrero. Understanding the dispersal of these languages over time is important for the prehistory of the region and for Mesoamerica as a whole. Previous research on Mesoamerican linguistic and cultural history is heavily focused on Mayan and Nahua peoples and their languages. Mixtecan peoples have received comparatively little attention, even though they once inhabited an even larger territory than today (Pérez Rodríguez 2013, Joyce 2011). In historical linguistics, the situation is not as dire, but longstanding preconceptions about the language family have influenced research on its reconstruction and dispersal and the internal structure and development of Mixtecan is still poorly understood. In a small-scale study on relatedness of seven Mixtec varieties spoken in the Juxtlahuaca district, Padgett (2017) found that current groupings in Ethnologue (Lewis et al. 2015) and INALI (INALI 2009a) do not align with speakers’ reports of mutual intelligibility. This has dire consequences when community materials are created on the basis of ‘established’ classifications like Ethnologue and Glottolog (Hammarström et al. 2021), which suffer from similar issues, and then are distributed to communities who cannot understand them. Such issues can be traced back to the characterization as dialect continuum and the comparative method and subgrouping might not be applicable here (Bradley & Josserand 1982:303).

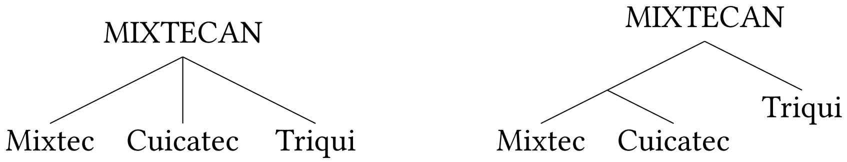
Bayesian phylogenetics have been applied to a wide array of language families (cf. Greenhill & Gray 2009, Lee & Hasegawa 2011, Kolipakam et al. 2018, Bouckaert et al. 2018, among others), but the only family studied with these methods (partially) located in Mesoamerica is Uto-Aztecan (Dunn et al. 2011). In the next section, we summarize earlier work on the classification of Mixtecan.

2.2 Previous Classifications and Dating Estimates

Longacre (1957, 1961) includes Triqui, Cuicatec, and Mixtec in the Mixtecan group of Otomanguean, with Amuzgo in a position coordinate to Mixtecan. Swadesh (1960) places Trique outside of Mixtecan and includes Amuzgo within it. In later work, Longacre (1966) reasserts that Amuzgo is not Mixtecan and expresses doubts about it even being Mixtecan’s closest relation within Otomanguean. Rensch (1976) adopts Longacre’s revised proposal, but more recent work by Kaufman (2006) agrees with Longacre’s original proposal, classifying Amuzgo as external to, but coordinate with, Mixtecan (see also Campbell 1997 and Campbell 2017b). Following the most recent consensus, with Amuzgo as a sister to Mixtecan, we exclude Amuzgo from our analysis.

Proto-Mixtecan forms have been reconstructed by various scholars (Longacre 1957, Gudschinsky 1959, Arana Osnaya 1960, Longacre 1962, Swadesh 1967), and these must be considered through the lens of which classification the authors proposed or supported at the time. Throughout the body of work on Mixtecan comparative studies, the internal subgrouping of the family remains poorly understood. Further disagreements pertain to whether Mixtec and Cuicatec form a branch vis-a-vis Triqui as Belmar (1902:4) suggests, or whether all three are coordinate with one another, cf. Figure 2.1. Macaulay (1996:6) states that the Mixtec-Cuicatec group has been convincingly demonstrated by Kaufman (1983, 1988), but both of these unpublished studies lack sufficient primary data for assessing the claim. The latter proposal of a flat structure is based on negative evidence, i.e. in the absence of convincing evidence for grouping Mixtec and Cuicatec closer together (Josserand 1983:101). In the following, we summarize the proposals of internal grouping for each of the three branches.

Triqui consists of three distinct varieties (Hollenbach 1977, DiCanio 2008, Matsukawa 2008) centered around the towns of San Andrés Chicahuaxtla, San Juan Copala, and San Martín Itunyoso, all in the state of Oaxaca (Hammarström et al. 2021).



(a) Josserand 1983:101

(b) Kaufman 1988

Figure 2.1: Two proposals for subgrouping within Mixtecan

Cuicatec is considered to consist of a handful of mutually-intelligible varieties (Anderson & Roque 1983), but unfortunately this group remains very sparsely documented (Campbell 2017a:8), and we will not be able to address its internal classification further.

Mixtec is by far the largest and most diversified Mixtecan subgroup and the languages are spoken in at least 189 municipalities in Oaxaca, Guerrero, and Puebla states in Mexico (Smith Stark 1994). It is characterized as a dialect continuum for which it is difficult or impossible to know how many languages and varieties there are. Jiménez Moreno (1962) considers Mixtec a ‘language’ with seven ‘dialect complexes’, but he does not provide primary data and argumentation. Hammarström et al. (2021) indicate that Mixtec comprises at least 53 varieties, but this appears to be based on Ethnologue, which in turn reflects the inter-intelligibility studies by Egland (1983). INALI (2009b) lists 81 varieties, with no supporting evidence for the determination of the groups and their membership. Such confusion proliferates in much of the literature on Mixtec languages and is only exacerbated by the unfortunate practice of referring to the Mixtec group as one language.[2]

The most systematic Mixtec subgrouping proposal to date is Josserand’s (1983) dissertation, in which she proposes 12 dialect areas (some with further subdivisions), reproduced in Figure 2.2. Her work is based on a word list with 188 items collected from 120 Mixtec speaking villages and remains the most comprehensive comparative Mixtec study to date. She focuses on sound changes in vowels but mentions consonantal developments as well. Tone is ex-

[2] For example in Macaulay (1996:1)’s grammar where the introduction states that “Mixtec is an Otomanguean language ”, only to clarify later on that “ ‘Mixtec’ really should be considered a group of related but distinct languages” (Macaulay 1996:6).

cluded since at that time there was limited availability of reliable material. She established the dialect areas based on bundles of isoglosses, which reflect vowel changes. It is important to point out that Josserand’s groupings do not directly translate into a family tree, nor were they intended to.

The tree provided in Glottolog (Hammarström et al. 2021) largely follows this outline, but also lists Amoltepec as an isolated primary daughter of proto-Mixtec. This variety is not included in Josserand’s (1983) dialect areas, listed as the source of the tree in Glottolog, but it is proposed as its own group in Bradley & Josserand (1982). Lower-level groupings do not follow Josserand (1983) in most cases; we infer that they follow Ethnologue (Eberhard et al. 2021). The same is true for the distinction between languages and dialects, which is not made in Josserand (1983) – nor in the present study, cf. Section 2.3. To sum up, there are many uncertainties with regards to sub-grouping among and within each of the three Mixtecan groups, especially as regards Mixtec.

Some of the earlier works mentioned above not only probed sub-grouping and reconstruction, but also provided dating estimates based on lexicostatistics and glottochronology (Holland 1959, Arana Osnaya 1960, Swadesh 1967). Both of these methods have severe shortcomings. Lexicostatistics is based on superficial similarity of form, without systematic cognate identification, and fails to account for situations in which lexical borrowing is extensive (Comrie 2000), and glottochronology makes the false assumption that lexical replacement occurs at a uniform, constant rate in all languages (Hoijer 1956, Nettle 1999, Holman 2010, among many others). These estimates, however, are the only ones currently available and therefore have been integrated into archaeological and anthropological studies (cf. Byers 1967, Flannery & Marcus 1983, Josserand et al. 1984). Mixtec and Triqui are estimated to have diverged around 3900-3500 years ago, Triqui and Cuicatec around 3900 years ago, and Mixtec and Cuicatec around 3100-2500 years ago (Holland 1959:25-26, Arana Osnaya 1960:262-263, Swadesh 1967:94).

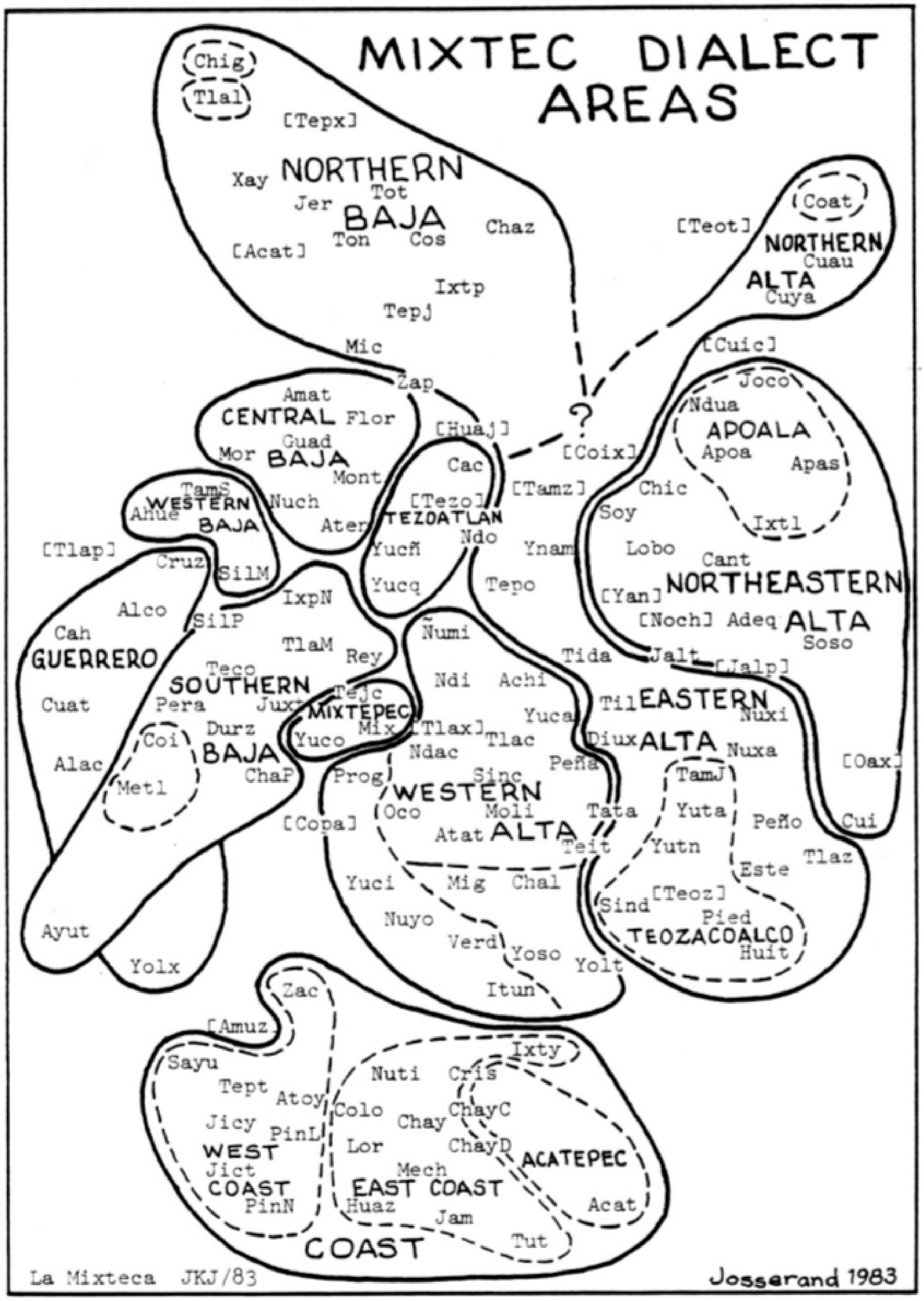


Figure 2.2: Mixtec Dialect Areas according to Josserand 1983:470

2.3 Data

2.3.1 Languages

We included every Triqui, Cuicatec, and Mixtec variety in the sample for which there is enough documentation and the materials are available to us. We do not distinguish between languages and dialects, since there is no solid basis nor necessary or sufficient criteria to do so in Mixtecan. 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 languages (Cruz & Woodbury 2014).

We collected data from 4 Triqui varieties, 3 Cuicatec varieties, and 130 Mixtec varieties, i.e. 137 languages in total. The Triqui varieties include the modern varieties mentioned in Section 2.2 – San Martín Itunyoso, San Juan Copala, San Andrés Chicahuaxtla – and the historical account of Chicahuaxtla Triqui spoken at the end of the 19th century (Belmar 1897). For Cuicatec, we include the closely related varieties spoken around the town of Santa María Pápalo, and the historical account of Tepeuxila Cuicatec spoken around the year 1900 (Belmar 1902). The Mixtec varieties are the 120 sampled in Josserand (1983), excluding proto-Mixtec. We added 10 varieties based on our own or colleagues’ documentation work and recently published sources. Of the varieties sampled in Josserand (1983), 17 were partially updated and provided with tone marking in Dürr (1987) and for another 10 there are new materials available. An overview of all languages and sources can be found in the Appendix in Table A.1 and a geographical overview in Figure 2.3.

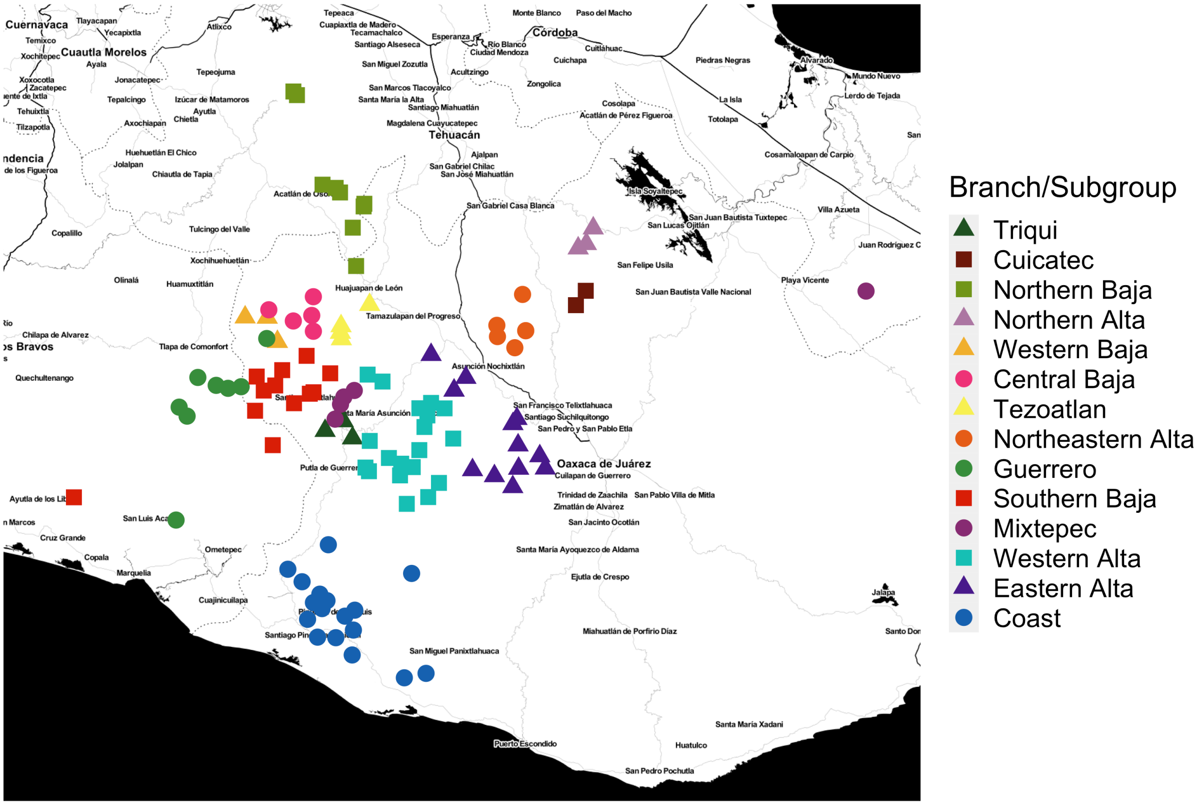


Figure 2.3: Sampled languages, with colors/shapes indicating primary branches and Josserand's dialect areas of Mixtec

2.3.2 Word list and data collection

We constructed a list of 209 concepts of basic vocabulary selected through a triangulation of already existing lists (with number of items in brackets: Josserand 1983 [188], Dürr 1987 [110], Padgett 2017 [98], Swanton & Mendoza Ruíz 2021 [86], and Campbell unpublished, [340]) and considerations about ease of elicitation (Laycock 1970) and semantics (Kassian et al. 2010). We kept all entries that appear in three or more lists and added a few that appear in two lists, but are easy to elicit and encode important cultural concepts. We excluded verbs because they require aspect-mood inflection, which is not well understood on a comparative Mixtecan level and is not always provided or reliably identifiable in glosses in the sources. The entire list is provided in SM 1.

We collected as many list entries as possible for each variety in the orthography provided in the source. We then converted the orthographic entries to IPA using orthography profiles (one for each source), which also standardize tone notation. The details of the conversion and standardization are laid out in SM 2. After data collection, we removed all concepts for which there was only one entry. Excluding duplicates, this results in a data set with 18,060 entries. The number of entries per variety ranges from 223 to 65, with mean at 131 and median at 125. We can thus say that on average we were able to gather around 60% of the concepts for each variety.

2.3.3 Cognate coding

The cognate coding was carried out based on the comparative method and mainly done by hand. We initially applied the 'LexStat' algorithm for automatic cognate detection (as implemented in List et al. 2019 and described in detail in List et al. 2018), but then refined and adjusted each cognate set by hand based on already established sound changes and correspondences (cf. Section 2.2 references), and our own knowledge of the languages in question. As

Mixtecan 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). This is exemplified with the entry for ‘scorpion’ in the Mixtec and Cuicatec varieties in Table 2.1, where all of the languages show an animate marker and the word for ‘tail’ for this concept. Since these are both recurrent morphemes with a clear meaning assigned to them, they each get their own cognate identifier (9 for the animate marker and 635 for the ‘tail’ morpheme). In Triqui, the entry for ‘scorpion’ is not further analyzable (and not cognate with ‘tail’) and so only has one cognate identifier assigned to it (749).

A novel feature we introduce is the annotation of tonal derivation. Mixtecan languages have highly complex systems of lexical and grammatical tone and the languages cannot be described or analyzed without making reference to tonal phenomena. 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. The only difference to segmental derivation is that the toneme cannot be visually segmented in the same way a concatenated affix may be. In Table 2.1, tonal derivation is exemplified by Alacatlatzala Mixtec in the word for ‘scorpion’. The second morpheme is clearly cognate with the word for ‘tail’, but in the derived word for ‘scorpion’, the second tone is raised. This tonal process is recurrent and represented with its own cognate identifier (1044). Note that in San Martín Duraznos Mixtec, we do not know whether this tonal process applied or not, since the base morpheme ‘tail’ was replaced by another (non-cognate) word. Consequently, we do not annotate it as being tonally derived.

We coded cognate sets in two ways: a broad analysis and a more fine-grained one. The fine-grained analysis takes into account potentially ‘irregular’ or ‘unexpected‘ reflexes, while the broad one ignores those. The reason for these two coding schemes is that these variant reflexes could carry phylogenetic signal that reflects shared innovations rather than parallel innovations. This could therefore be useful for sub-grouping. To be able to address this

Table 2.1: An example of the annotation of partial cognates and tonal derivation (Forms are given in IPA. In these examples, there is no difference between the broad and fine-grained cognate assignments.)

DOCULECT	CONCEPT	FORM	COGNATE IDs
AlacatlatzalaMixtec	TAIL	s i ^{1} ʔ m a ^{1}	635
MagdalenaPenascoMixtec	TAIL	s u ^{3} ʔ m a ^{1}	635
SanMartinDuraznosMixtec	TAIL	^{n}d o ^{3} ʔ o ^{1}	748
SantaMariaPapaloCuicatec	TAIL	^{n}d u ^{4} k u ^{1} + ð e ^{1} ʔ ẽ ^{3}	709 635
SanMartinItunyosoTriqui	TAIL	t u ^{3} n e ʔ ^{3}	749
AlacatlatzalaMixtec	SCORPION	t i ^{5} + s i ^{1} ʔ m a ^{3}	9 1044 635
MagdalenaPenascoMixtec	SCORPION	t i ^{1} + s u ^{3} ʔ m a ^{1}	9 635
SanMartinDuraznosMixtec	SCORPION	t ɕ i ^{1} + s u ^{3} ʔ m a ^{1}	9 635
SantaMariaPapaloCuicatec	SCORPION	i ^{3} t + ð e ^{3} ʔ ẽ ^{1}	9 635
SanMartinItunyosoTriqui	SCORPION	tʃ i ^{3} k ĩ ^{32}	636

question empirically rather than prejudging the matter, we opted for annotating the cognate sets in two ways. This allows us to run the models on both sets and assess the influence (or absence thereof) on the outcomes. The different codings are illustrated in Table 2.2. In both the reflexes for ‘bird’ and ‘frog’ in Table 2.2, some Mixtec varieties show an initial [l] while others have an [s] in this position. This correspondence surfaces in a few other items, but is not regular in the strict sense. Furthermore, it is unclear whether they represent different reflexes of the same proto-form or whether one or both of them contain remnants of fossilized prefixes. In the broad cognate coding, this difference is glossed over, while in the fine-grained one the [l]-forms are grouped into a separate class from the [s]-forms.

The assessment results in 1120 cognate sets in the broad analysis and 1197 in the fine-grained one. Details regarding cognate coding and the the full database of cognate sets can be found in SM 2 and SM 3.

This cognate database was converted to a binarized cognate matrix to serve as the input for the Bayesian phylogenetic analysis. This cognate matrix has one column for each cognate set, which marks the presence (1), absence (0), or lack of information (?) of this cognate set

Table 2.2: An example of the annotation of broad vs. fine cognate IDs (Forms are given in IPA)

DOCULECT	CONCEPT	FORM	BROAD	FINE
SanAndresYutatioMixtec	BIRD	l a ^{3} a ^{3}	49	49L
SantaMariaPenolesMixtec	BIRD	t ɨ ^{5} + l a ^{1} a ^{5}	9 49	9 49L
SantaMariaJicaltepecMixtec	BIRD	s a ^{3} a ^{3}	49	49
SanMiguelElGrandeMixtec	BIRD	t ɨ ^{3} + s a ^{3} a ^{1}	9 49	9 49
SanAndresChicahuaxtlaTriqui	BIRD	ʃ a ^{3} + t a^{h} ^{32}	9 49	9 49
SanAndresYutatioMixtec	FROG	l a ^{5} ʔ o ^{1}	263	263L
SantaMariaPenolesMixtec	FROG	l a ^{1} ʔ β a ^{5}	263	263L
SantaMariaJicaltepecMixtec	FROG	s a ^{3} ʔ β a ^{3}	263	263
SanMiguelElGrandeMixtec	FROG	s a ^{3} ʔ β a ^{1}	263	263
SanAndresChicahuaxtlaTriqui	FROG	ʃ i ^{2} + r i ^{3} k ɨ^{h} ^{3}	9 266	9 266

for each variety. We excluded 27 varieties due to low coverage (marked by * in Table A.1). This coding resulted in a cognate matrix with 1183 states (columns) for the broad cognate assignment and 1254 states (columns) for the fine-grained cognate assessment. The resulting data files in Nexus format (Maddison et al. 1997) are provided SM 4.

2.4 Methods

2.4.1 Tree-likeness and conflicting signal

To visualize and quantify the conflicting signal in the data set, we calculated δ-scores and Q-residuals. Both of these metrics assess how much conflicting signal there is in the data, or in other words, how tree-like the data is (Bryant & Moulton 2004). The Q-residuals and δ-scores were calculated with SplitsTree4 (Huson & Bryant 2006), which was also used to produce NeighborNets (provided in SM 5). δ-scores are computed for each tip and result in a number between 0 and 1, with higher numbers indicating more conflicting signal. Gray et al. (2010:3925) argue Q-residuals are a more direct measure of how much the tips diverge from

Table 2.3: Mean Delta-scores and Q-residuals

Cognate Sets/Family	δ-score	Q-residual	Source
broad	0.3911	0.02554	
fine	0.3793	0.02472	
Polynesian	0.41	0.002	Gray et al. 2010
Dravidian	0.30	0.0069	Kolipakam et al. 2018
Chapacuran	0.262	0.016	Birchall et al. 2016

a strict tree, because they take into account potential effects of scaling. A lower Q-residual score reflects more adherence to a strict tree. Both scores can be averaged over all tips to give a measure of tree-likeness of the whole network. These average scores are summarized in Table 2.3. The scores for each tip are provided in SM 2. The broad cognate sets result in slightly higher scores as compared to the fine-grained sets, exhibiting more conflicting signal or, in other words, a less tree-like structure. The potentially 'irregular' reflexes annotated in the fine-grained sets should thus be investigated in more detail when family-wide sound changes are worked out, since they are potentially useful for sub-grouping.

Delta-scores and Q-residuals cannot be straightforwardly compared across data sets of different languages, and there are no clear cut-offs of what counts as tree-like or not (Gray et al. 2010:3925-3926). It is still instructive to situate our results in the context of other studies, given that the Mixtec group has been explicitly described as evolving in a non-tree-like manner (cf. Section 2.2). The scores of three other language families are also summarized in Table 2.3. Mixtecan has a δ-score almost as high as Polynesian, which is described as very reticulate, and a Q-residual score higher than all three language families. This corroborates the impressionistic views that Mixtecan languages exhibit a high degree of conflicting signal.

The amount of reticulation, however, is not distributed evenly among Mixtecan languages. To better illustrate this, we group the languages according to Josserand's proposed dialect areas and plot the Q-residuals and δ-score against each other in Figure 2.4. First, we note

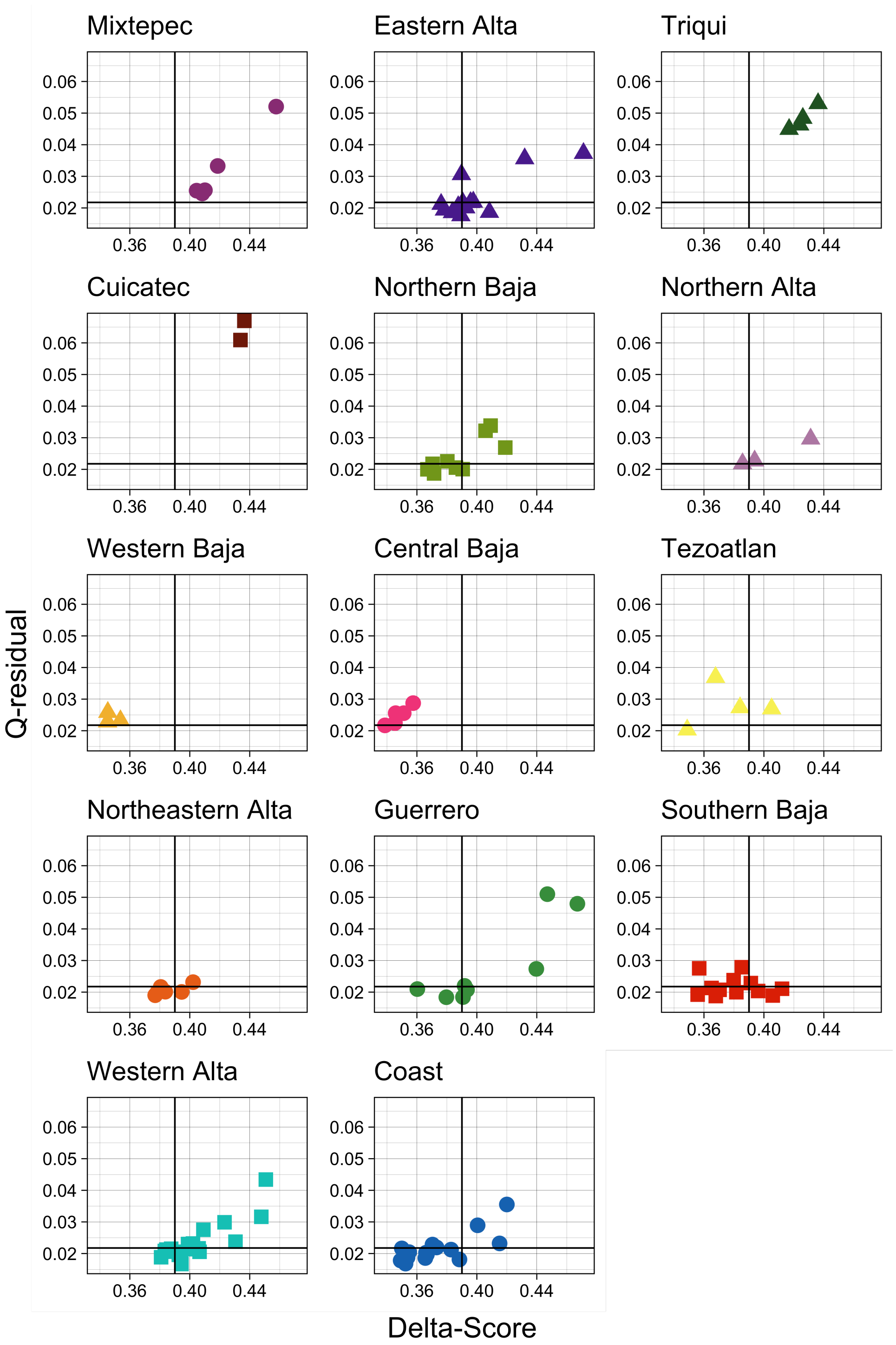


Figure 2.4: Delta-scores and Q-residuals plotted against each other by dialect area according to Josserand (1983)

that Cuicatec and Triqui languages show high scores – in fact higher than most of the Mixtec languages. This is not surprising with regards to Cuicatec. This branch is the least well studied of the three, with little in-depth documentation and almost no historical work on those languages. It is thus possible that there are undetected loans from Mixtec languages in the data.

The amount of conflicting signal in Triqui languages requires another explanation. The Triqui people are a quite small group in the region and their territory is completely surrounded by and in some cases shared with Mixtec speakers. The Triqui communities have long had a high level of contact with Mixtec speakers. Historically, Triqui speakers had some degree of bilingualism with San Juan Mixtepec Mixtec (a nearby Mixtec language) and they travelled regularly to Mixtec towns (especially Cuquila and Tlaxiaco) for the purposes of commerce (see DiCanio 2022a). This could help explain their non-tree-like history.

The Mixtec groups with the lowest amount of conflicting signal are the Central and Western Baja, the overall highest scores are found in the Mixtepec group. Very high scores – or a lot of conflicting signal – are also found in a few Guerrero languages and some languages of the Western Alta.

To explore these data further, we apply a Bayesian phylogenetic model. Bayesian phylogenetic methods are well suited to explore language history based on cognate data and have several advantages over other methods. Unlike lexicostatistics, they allow for rate-of-change variation across languages, across cognates, and over time. Bayesian phylogenetic methods do not produce a single ‘best’ tree, but rather sample the space of possible trees to return a distribution of trees that fit the data well given the model. This posterior sample provides a natural way for calculating the support for particular groupings while allowing us to also take into account differing scenarios. This means that we can quantify uncertainty in the tree, both with respect to nodes (or splits) and with respect to model parameters. Furthermore, Bayesian phylogenetic approaches incorporate linguistic and historical knowledge into the model via priors and calibration points. For a detailed explanation of Bayesian phylogenetic methods see Hoffmann et al. 2021.

2.4.2 Calibration Points

To infer dates, calibration points are needed. Unfortunately, the Mixteca region has not received much attention by archaeologists and anthropologists, but the area around Monte Albán and the Mixteca Alta are fairly well excavated (Pérez Rodríguez 2013). Although recent years have seen many advances, it is still difficult to identify specific cultural groups in the archaeological record (Pérez Rodríguez 2013:93), especially in early periods before people adopted a mostly sedentary lifestyle.

The Mixtecs have chronicled their royal lineages in several codices (cf. Jansen 1990 for an overview). Those that survive to the present day are from the late post-classic and early colonial period. These are pictorial manuscripts and, while clearly identifiable as Mixtec, we cannot tell with certainty which variety of Mixtec was spoken by their creators or by the people listed in the codices (Jansen & Pérez Jiménez 2011:7).

This lack of documentation leaves us with few candidates for calibration points. There are historical documents of the early and late colonial period (de los Reyes 1890 [1593], de Alvarado 1962 [1593], Belmar 1897, 1902) with vocabularies and grammar sketches that allow us to clearly identify the language described. The Mixtec variety of Teposcolula was used as a *lingua franca* throughout the 16th century (Jansen & Pérez Jiménez 2011:9) and was thus documented quite extensively by the Dominicans. Given that we know when these documents were published, we can use that date as a prior setting for the Teposcolula Mixtec variety. In the late colonial period, Belmar documented a variety of Cuicatec and a variety of Triqui (Belmar 1897, 1902). These two documents also serve as calibration points, all summarized in

Table 2.4: Calibration Points

Language(s) affected	Type	Setting	Details
Teposcolula Mixtec	tip date	350 BP	age of historical document
Tepeuxila Cuicatec	tip date	50 BP	age of historical document
Chicahuaxtla Triqui	tip date	60 BP	age of historical document

Table 2.4.[3]

2.4.3 Model Specifications and Comparison

We carried out a Bayesian phylogenetic analysis with BEAST2 (Bouckaert et al. 2019). BEAST2 estimates the posterior distribution of trees using a Markov Chain Monte Carlo (MCMC). As the tree prior, we used a Birth-Death Skyline with Serial Sample model parameterized on birth, death, and sampling rates (Stadler et al. 2013). This tree prior controls how the sampled trees are initially built. Under the Birth-Death Skyline model lineages are born (= ‘birth’) and go extinct (= ‘death’) at separate rates. We parameterized the birth and death rate as exponentially distributed with a mean of 0.01. This means that, on average, a new lineage is born every 100 years, while an existing lineage lives for an average of 100 years (Hoffmann et al. 2021).

The third parameter in this model is the sampling parameter. A language is said to be ‘sampled’ if it is included in the data set, but there is an unknown number of languages that existed in this family through time that fell out of use before being recorded. However, it is unlikely that Mixtec was as highly diversified as it currently is before or even at the time of the Spanish conquest, and we thus estimate that we have been able to sample more than half of the languages. We modelled this parameter following a Beta distribution with a mean of around 60% (Beta[110, 80], cf. Hoffmann et al. 2021).

[3] As is common in phylogenetic studies, the ‘present’ is set at 1950 like in archaeology, so we can cite it as ‘Before Present’.

To model how the cognates change we applied a binary covarion model of cognate evolution (Penny et al. 2001). The covarion model estimates at which rate cognates are lost and gained over time and allows for the rates to vary, such that there can be 'slow' and 'fast' periods of changes in cognate sets. This accounts for the intuition shared by many linguists that some cognates can be relatively stable over a period of time and then change rather rapidly (e.g. due to language contact). To model rate variation over time we fitted both a strict clock and a relaxed clock model of character change (Drummond et al. 2006). The strict clock assumes the same rate of substitution across the whole tree, while the relaxed clock allows the rates to vary across branches.

This setup resulted in a total of four models, summarized in Table 2.5. The Maximum Clade Credibility trees was extracted with TreeAnnotator (part of the BEAST2 distribution) and the graphical representations created with the package ggtree (Yu et al. 2017) in R R Core Team (2021). The visual representation of all posterior trees was obtained with DensiTree (Bouckaert & Heled 2014). The BEAST XML files and the full MCC tree are available in SM 6 and 7.

To evaluate which model fits these data best, we used nested sampling to calculate the marginal likelihoods of each model (Maturana Russel et al. 2019) and Bayes Factors to quantify the differing model performance (Kass & Raftery 1995). We ran each model for 100,000,000 generations, sampling every 5000th generation to avoid auto-correlation. We discarded the first 10% as 'burn-in' where the inferred parameters were still unduly affected by the Markov Chain's starting parameters. The trace files were inspected with Tracer 1.7 (Rambaut et al. 2018), which showed that all the models converged after burn-in and critical parameters all showed effective sample sizes above 200. Table 2.5 summarizes the results of the model comparison. The best performing model, i.e. the one with the highest Bayes Factor, is the one based on the broad cognate assignments with a relaxed clock. This model outperforms the strict clock model with the same data set. This is true also of the two models based on the fine-grained cognate assignments: the relaxed clock model performs better than the strict clock one. However, both models based on the fine sets have lower Bayes Factors than the ones based on the broader cognate sets.

Table 2.5: Model Comparison results sorted by Bayes Factor[4]

Cognate Coding	Clock Model	MlogL	logBF	SD
broad sets	relaxed clock	-11325		3.72 **
broad sets	strict clock	-11469	144	3.23
fine sets	relaxed clock	-12253	784	3.49 **
fine sets	strict clock	-12406	153	3.20

As our analysis only contained minimal calibrations we were concerned there would not be enough signal to provide a robust estimate of the age of the family. Therefore, we formally tested whether the temporal information provided sufficient signal using a Bayesian estimation of temporal signal (“BETS”) analysis (Duchene et al. 2020) – i.e. we ran each analysis a second time without any temporal information and compared the model fit with and without temporal information. Table 2.6 provides the results of the BETS analysis. Overall, the temporal information has little impact on our models – as might have been expected given that all we have are three tip dates (cf. Section 2.4.2). In the models based on the fine sets the temporal information neither improves nor renders the analysis worse, since the Bayes Factors overlap. In the model with the broad cognate sets and a strict clock, the timed analysis outperforms the untimed one, but the reverse is true for the relaxed clock model with the broad sets. However, in both cases the practical difference is marginal and the tree topology is almost the same, except that the timed analysis shows overall lower posteriors for most of the nodes. We thus base the remainder of this paper on the model based on the broad cognate sets with relaxed clock and its output.

[4] Abbreviations used in the tables: MlogL = Marginal log Likelihood, logBF = log Bayes Factor, SD = Standard Deviation

[5] For abbreviations see footnote 4.

Table 2.6: BETS (Bayesian Estimation of Temporal Signal) results[5]

Cognate Coding	Clock Model	Temp. Inf.	MlogL	logBF	SD
broad sets	relaxed clock	+TIME	-11325		3.72
broad sets	relaxed clock	-TIME	-11311	14	3.47 **
broad sets	strict clock	+TIME	-11469		3.23 **
broad sets	strict clock	-TIME	-11460	9	3.08
fine sets	relaxed clock	+TIME	-12253		3.49
fine sets	relaxed clock	-TIME	-12248	5	3.50 * (overlaps)
fine sets	strict clock	+TIME	-12406		3.20 * (overlaps)
fine sets	strict clock	-TIME	-12409	3	3.20

2.5 Results

We discuss the results of the best performing model (broad cognate coding with relaxed clock) with respect to sub-grouping based on the Maximum Clade Credibility (MCC) tree and the densitree representations. We also refer back to the NeighborNet of the broad cognates sets where this is illustrative. We consider node posteriors of the MCC tree of over 0.7 as well supported and posteriors of over 0.9 as very reliable and will primarily discuss such nodes, starting from the root of the tree moving to lower level groupings. We refer to the groups identified in our model with numbers, using single digits for the higher level groupings and adding digits for each lower level. The numbers are intended as neutral labels. We did not use capital letters (as is customary for example in Bantu classification) to avoid confusion with Josserand’s (1983) previously established dialect areas and the geographic areas (Alta, Baja, and Costa) of the Mixteca region.

Our model recovers the three branches of Mixtecan – Mixtec, Cuicatec, Triqui – well, cf. Figure 2.5. Triqui and Cuicatec are grouped together versus Mixtec, even though we expect Mixtec and Cuicatec grouped together (cf. Section 2.2). This unexpected grouping was also present in the untimed model, suggesting that it is not the temporal information from the late

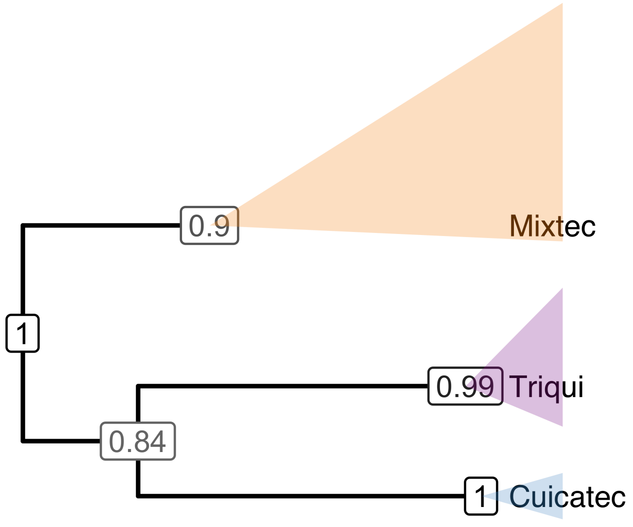


Figure 2.5: Primary branches of Mixtecan as recovered by our model with node posteriors

colonial Triqui and Cuicatec documents primarily responsible for this division. As mentioned in Section 2.2, both Cuicatec and Triqui have been influenced by Mixtec speakers, who are and have been the majority in the area. However, there were also some shared Triqui-Cuicatec cognates which lacked Mixtec language cognate forms and are at least partially responsible for their closeness in the tree and the NeighborNet. It should be further investigated which factors led to Triqui and Cuicatec being grouped together in our model.

Since only two varieties of Cuicatec could be included, one of which is not contemporary, the internal structure of Cuicatec cannot be discussed further. Within Triqui, we do not find well supported divisions, apart from the one that separates the colonial Chicahuaxtla variety from the contemporary ones. We thus refrain from any further interpretation of the internal groupings of Triqui.

In Mixtec, which comprises the largest number of varieties of the three branches, we discuss each well supported group and linkage in terms of earlier proposals and implications for further research. We use the term ‘linkage’ to refer to low-level groups and varieties which are placed close together but do not have good support for forming a group (Ross 1988, François 2015). An overview of the higher level groups and linkages is provided in Figure 2.6.

The first split in Mixtec separates the colonial era Teposcolula Mixtec from all other varieties, see Figure 2.7. Teposcolula was the political center during the colonial period and the variety spoken there served as a *lingua franca* throughout the Mixteca. Josserand (1983) classified Teposcolula as belonging to the Eastern Alta group. This is not recovered in our model, but it is possible that the temporal distance influenced the placement of this variety since it is about 350 years removed from contemporary varieties. In the untimed model, Teposcolula is part of Group 1 (which corresponds to Josserand's (1983) Northern Alta), albeit not with high certainty (posterior = 0.72). In the NeighborNet, this variety is placed between varieties of Josserand's Eastern Alta and Coast groups, but closer to the Eastern Alta. All of this suggest that further research is needed to clarify the relationship of the Teposcolula variety to contemporary Mixtec varieties.

The next well supported split sets Group 1 apart from all the other Mixtec groups and linkages, cf. Figure 2.8, and corresponds to Josserand's (1983) Northern Alta group. The towns where these varieties are spoken are geographically separated from other Mixtec speakers. They are completely surrounded by Mazatec (Otomanguean: Popolocan) speaking communities (Josserand 1983:104) and also border the Cuicatec region, as well as Ixcatec and Chocho speaking towns. The Mixtecs migrated to this area after the Mazatecan communities were already established (cf. Gudschinsky 1958 for the influence of Mixtec on Mazatec dialect history). These facts explain well why Group 1 is set apart from all others. The separation of Group 1 from the rest of the Mixtec varieties is also well reflected in the Densitree visualization and the NeighborNet (cf. SM 5 and 8).

Group 2 corresponds to the Coast group from Josserand (1983). It is clearly set apart from other groups in the Densitree visualization and also clusters together in the NeighborNet,

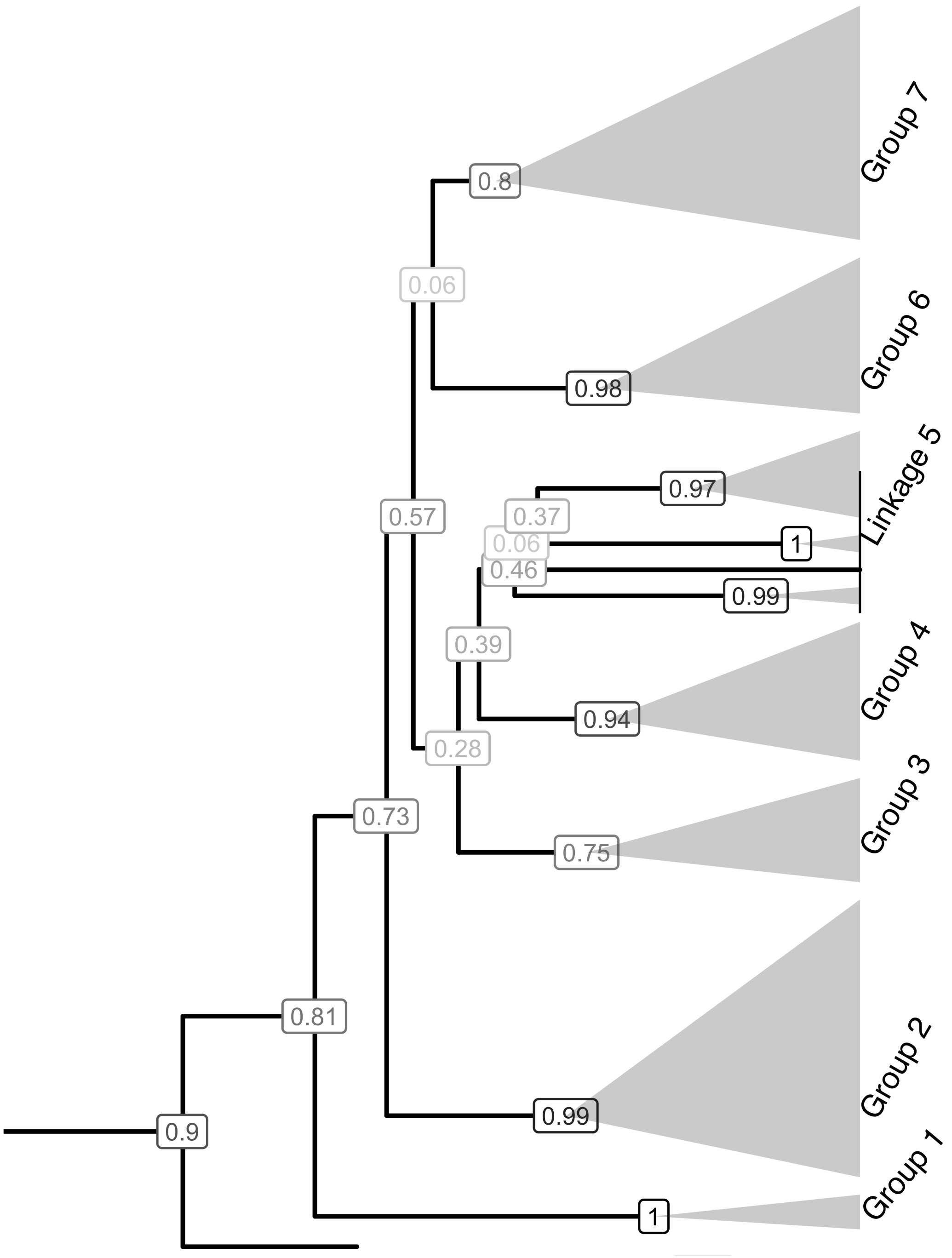


Figure 2.6: The Mixtec groups and linkages with node posteriors

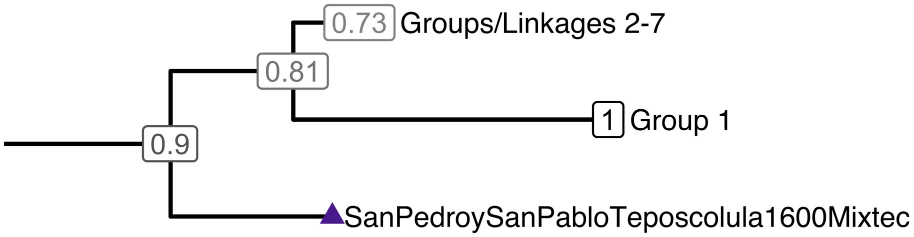


Figure 2.7: The first and second split within the Mixtec branch with node posteriors

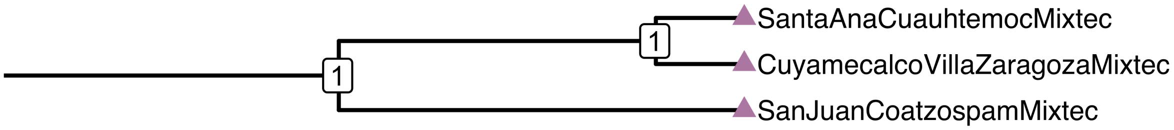


Figure 2.8: Mixtec Group 1 with subdivisions and node posteriors

showing less reticulation than other groups. The Mixtecs did not originally occupy territories on the coast, but rather emigrated there. Although these towns are connected to the Mixteca Alta by trade – both today and historically – they are more isolated from other Mixtec speakers than those of the other regions (Josserand 1983:116). They are, however, not isolated from each other, which explains the high degree of cohesion of this group in terms of cognacy data. This is reflected in the Q-residuals, which are relatively low for most coastal varieties (cf. Figure 2.4). There is a relatively well supported three-way partition within Group 2, cf. Figure 2.9. Subgroup 2.1 separates out Ixtayutla Mixtec (included in the East Coast group in Josserand 1983). Ixtayutla is relatively remote from other Mixtec towns and is described as conservative, located at the eastern boundary of the coastal Mixtec region (Josserand 1983:116) and some villages in that region are bilingual with Zenzontepec Chatino. This explains its separate position in the MCC tree well. Subgroup 2.2 broadly reflects Josserand’s (1983) West and East Coast groups. However, Subgroup 2.2.1, roughly corresponding to the East group, also contains the variety of Acatepec, singled out by Josserand as belonging to neither group. Subgroup 2.2 (roughly Josserand’s West group) includes the variety of San Juan Colorado, classified as belonging to the eastern group in Josserand (1983). This opens avenues for further research. Since Josserand’s classification is predominantly based on vowel correspondences, it is possible – and should be investigated – that considering sound changes of these varieties on a broader scale places them in the same groups as the lexical cognacy data.

Group 3 comprises a part of the varieties classified by Josserand (1983) as Western Alta. They are spoken in a roughly contiguous area around the Tlaxiaco district in the western part of the Mixteca Alta. Subgroup 3.1 sets the varieties of Teita and Yucuañe apart from the others, cf. 2.10. These varieties are also set apart from the others in the NeighborNet, where they appear closer to Linkage 5 and Group 6 varieties.

Group 4 broadly corresponds to Josserand’s (1983) Eastern Alta group, although excluding the Diuxi, Tilantongo, and Tidaa varieties which are part of Linkage 5 in our model, cf. Figure 2.11. These varieties are also set apart from the other ‘Eastern Alta’ ones in the NeighborNet, where they appear closest to Linkage 5 languages. As can be seen in Figure 2.6, the relationship of Linkage 5 to the other groups is rather unclear, with all the intermediate nodes exhibiting very low posteriors. It is thus possible that further research could provide evidence for Josserand’s (1983) original Eastern Alta group.

Linkage 5 consists of varieties and small groups that are linked together and connected to the other groups by nodes with low to very low posteriors. There is, however, one group within this linkage which has a high posterior and is well supported. This is Group 5.1, which corresponds to Josserand’s (1983) Northeastern Alta group. There are two well supported subdivisions, of which one sets the varieties of Apazco and Jocotipac apart from the others. This is interesting because the varieties of Apazco and Apoala are represented as dialects of one language in Glottolog (Hammarström et al. 2021), while Soyaltepec is set apart as its own language. However, Apoala is a very conservative Mixtec variety and surrounding towns in the same municipality show a lot of variation. In light of this and our results, there is no support for viewing Apoala and Apazco Mixtec merely as dialects of each other.

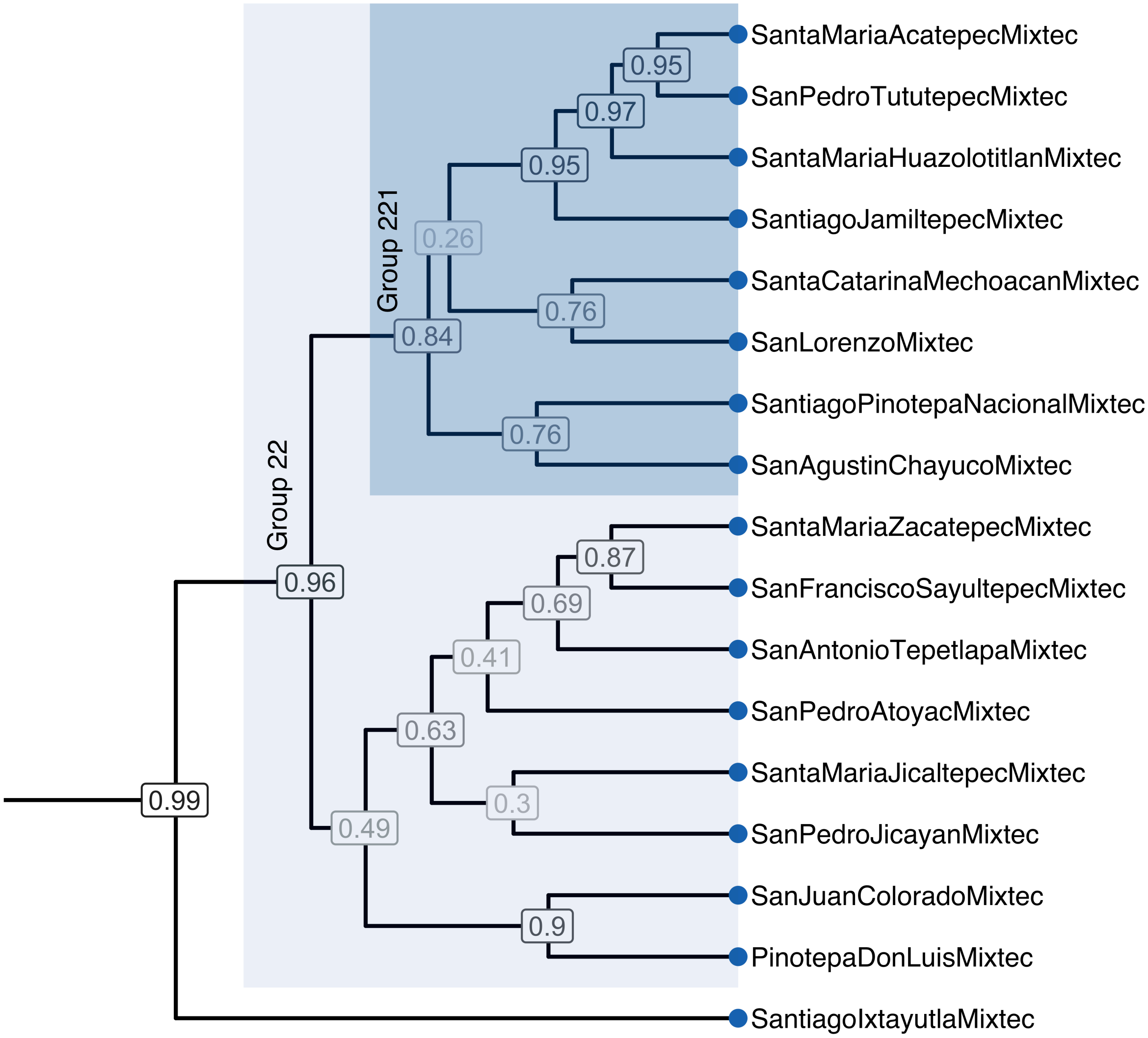


Figure 2.9: Mixtec Group 2 with subdivisions and node posteriors

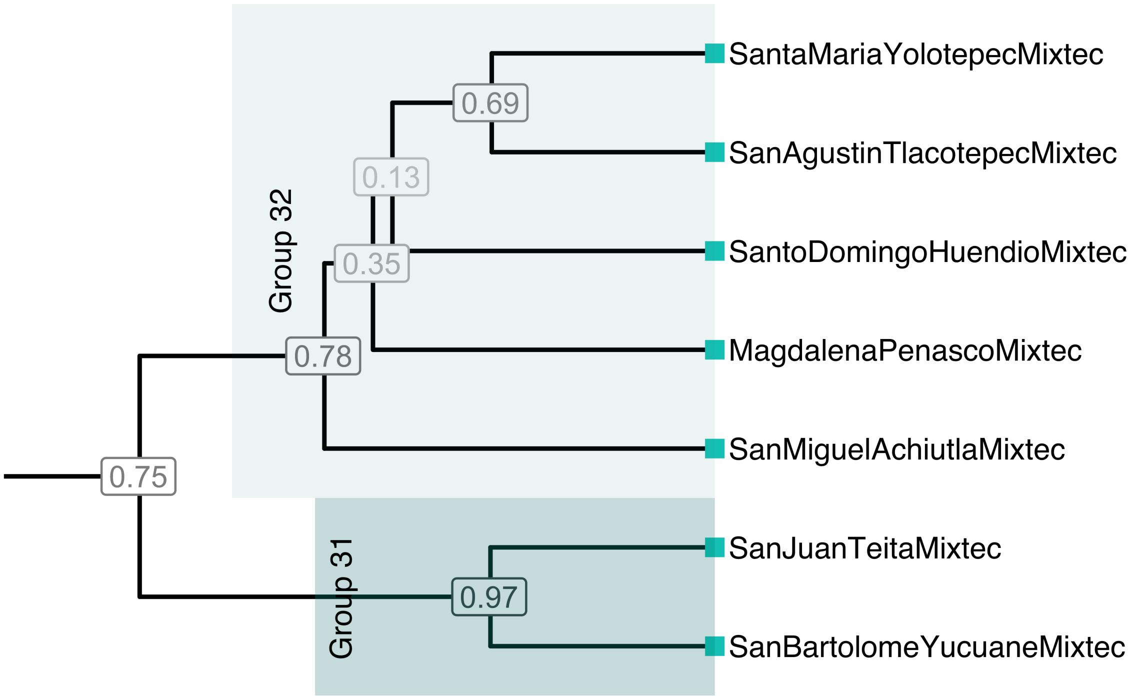


Figure 2.10: Mixtec Group 3 with subdivisions and node posteriors

Of the other varieties, those of Yosonama and Ñumi are closely related. In Josserand (1983) they are classified as Western Alta (our Groups 3 and 6). The varieties of Diuxi and Tilantongo are also grouped together closely. They and the not directly connected variety of Tidaa are classified in Josserand (1983) as Eastern Alta (our Group 4). As mentioned above, the posteriors connecting these varieties are very low and it thus remains an open question whether any of them form a closer relationship with each other or other groups outside of Linkage 5. These varieties also appear next to each other in the NeighborNet, but with a lot of reticulation between them.

Group 6 consists of varieties roughly situated in the western part of the Tlaxiaco district. In Josserand (1983), these varieties form part of the Western Alta group (together with our Group 3 and the varieties of Yosonama and Ñumi in Linkage 5). In the NeighborNet, they are placed close to Group 3 varieties, suggesting that further research could reveal a closer

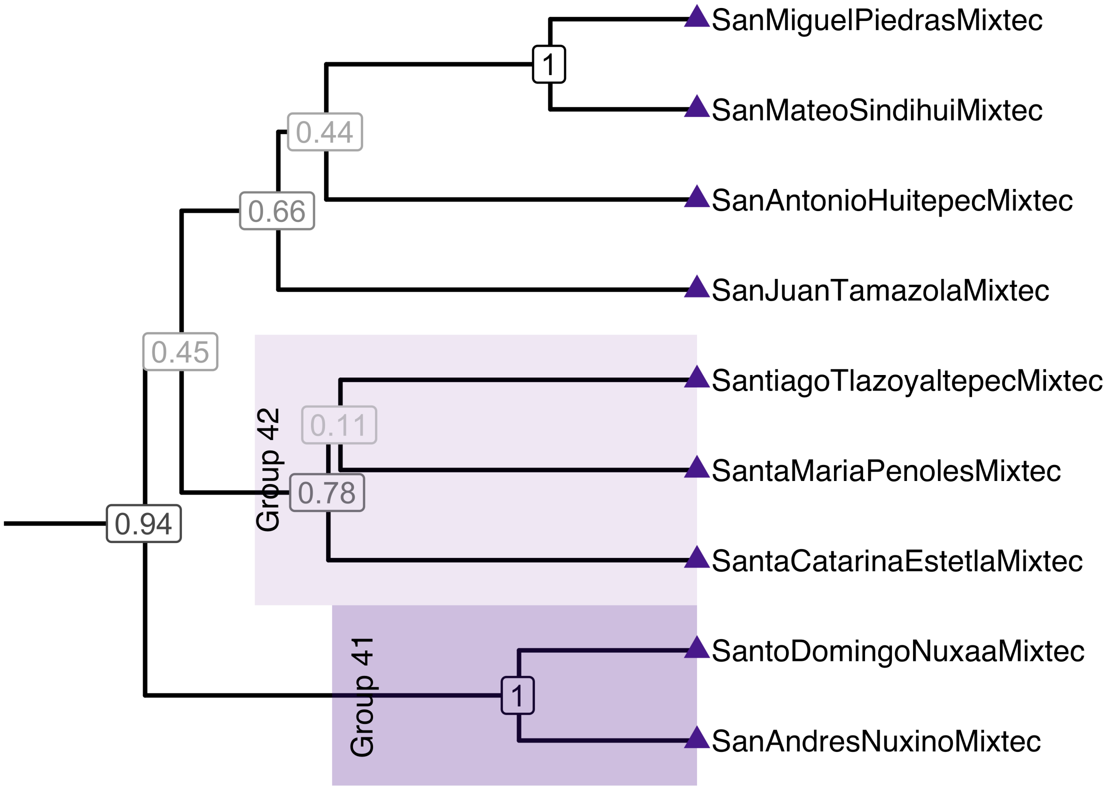


Figure 2.11: Mixtec Group 4 with subdivisions and node posteriors

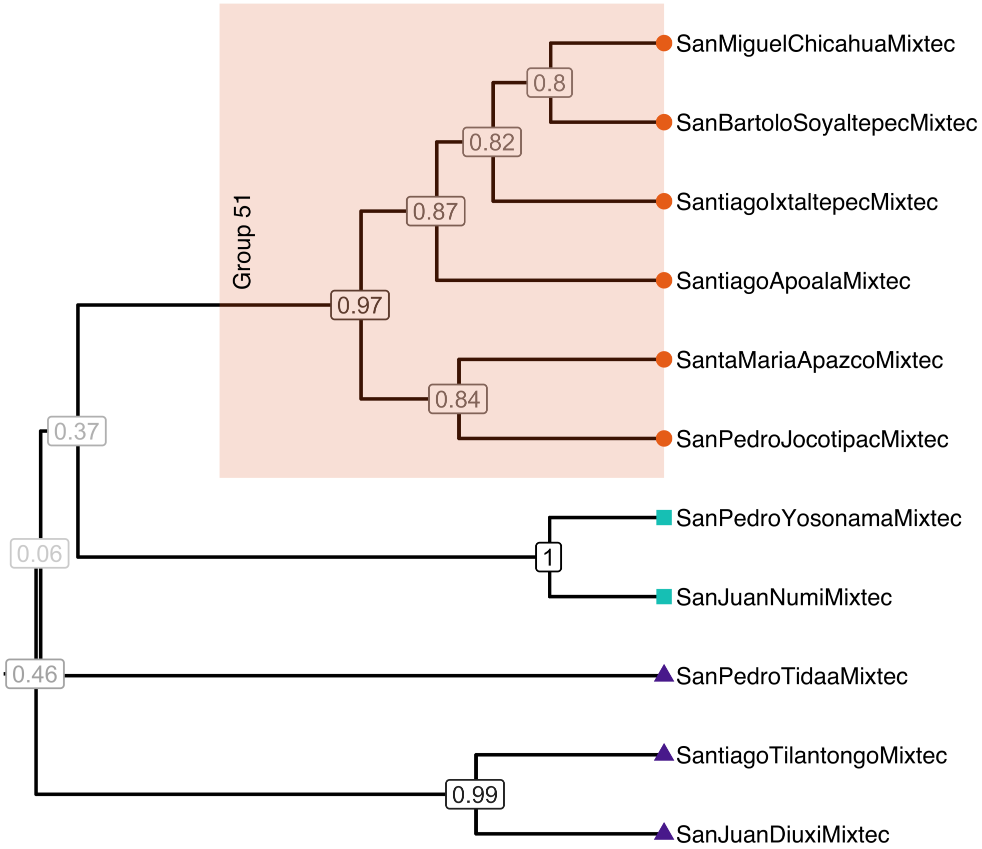


Figure 2.12: Mixtec Linkage 5 with subdivisions and node posteriors

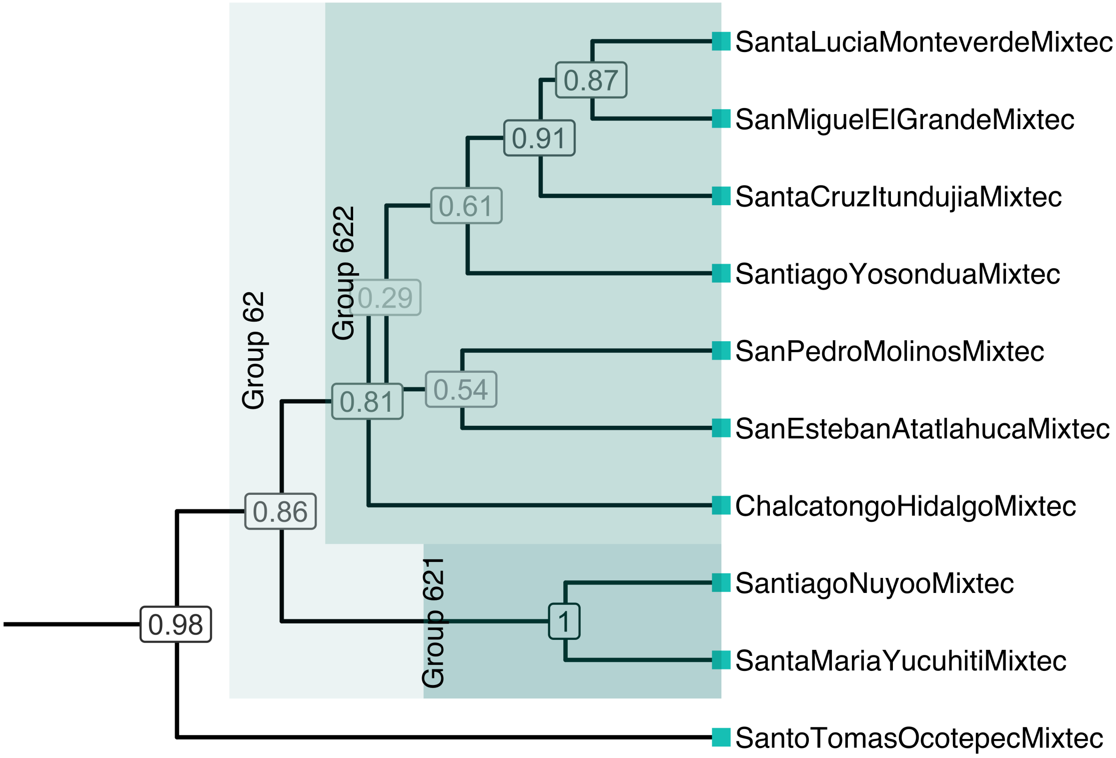


Figure 2.13: Mixtec Group 6 with subdivisions and node posteriors

relationship. Internally, the variety of Ocotepec is set apart from the others and a larger second grouping (Subgroup 6.2) consists of the rest of the varieties, cf. Figure 2.13. Within Subgroup 6.2, we find a smaller division (6.2.1) consisting of the varieties of Yucuhiti and Nuyoo set apart from the rest (6.2.2). The internal composition of this subgroup invites further research, as the varieties of Chalcatongo and Molinos are said to be very similar to that of San Miguel el Grande (they are represented as dialects of it in Hammarström et al. 2021), but are placed relatively far from it in our results.

The next larger grouping that is relatively well supported (Group 7 with posterior = 0.8) contains varieties spreading over 7 of the groups identified by Josserand: Northern Baja, Central Baja, Western Baja, Southern Baja, Guerrero, Mixtepec and Tezoatlan, cf. Figure 2.14. It it thus worth discussing its internal structure in more detail. In the NeighborNet, this group forms a large ‘fan’ complex apart from the other varieties with a large degree of reticulation. We identify four subgroups and two linkages, as well as two varieties not clearly placed in a group or linkage, cf. Figure 2.14. The higher level nodes – representing how these subgroups are connected to each other – have low posteriors, suggesting further research is needed in this area.

Subgroup 7.1 corresponds to Josserand’s (1983) Mixtepec group, cf. Figure 2.15. It shows a further partition separating the variety of Yucunicoco from the others. This is congruent with the fact that Yucunicoco is located further away from the other villages, closer to the Triqui area. The issue in need of discussion is the classification of the variety of Santiago Juxtlahuaca in this group rather than Linkage 7.2. This variety belongs to the Southern Baja group according to Josserand and should be most closely related to the variety of Tecomaxtlahuaca and surroundings (see below on linkage 7.2). However, earlier proposals such as Mak & Longacre 1960 and Bradley 1968 did place Juxtlahuaca with the Mixtepec varieties. There are several reasons why difficulties might arise in the classification of this variety. Santiago Juxtlahuaca is the capitol of the Juxtlahuaca district and also the seat of the municipality of Juxtlahuaca. As such, it is not only the biggest town in the area, it also sees a continuous and considerable influx of people from surrounding towns, both Mixtec and Triqui. This means that there are speakers of different varieties present (as well as many who only speak Spanish), which is likely to influence the local variety of this town, for which little data is available (only those collected in Josserand 1983).

Subgroup 7.3 covers all varieties from Josserand’s (1983) Guerrero group and a portion of her Southern Baja group. Linkage 7.2, also illustrated in Figure 2.16, is very small and contains only varieties of Josserand’s (1983) Southern Baja group. However, the two groups from Josserand are not clearly separated in our results, cf. Figure 2.16, and there appear to be a few ‘misplaced’ varieties, such as Tlahuapa Mixtec and Tepango Mixtec. The posterior of the node above Linkage 7.2 and Group 7.3 is very low, indicating uncertainty about a higher

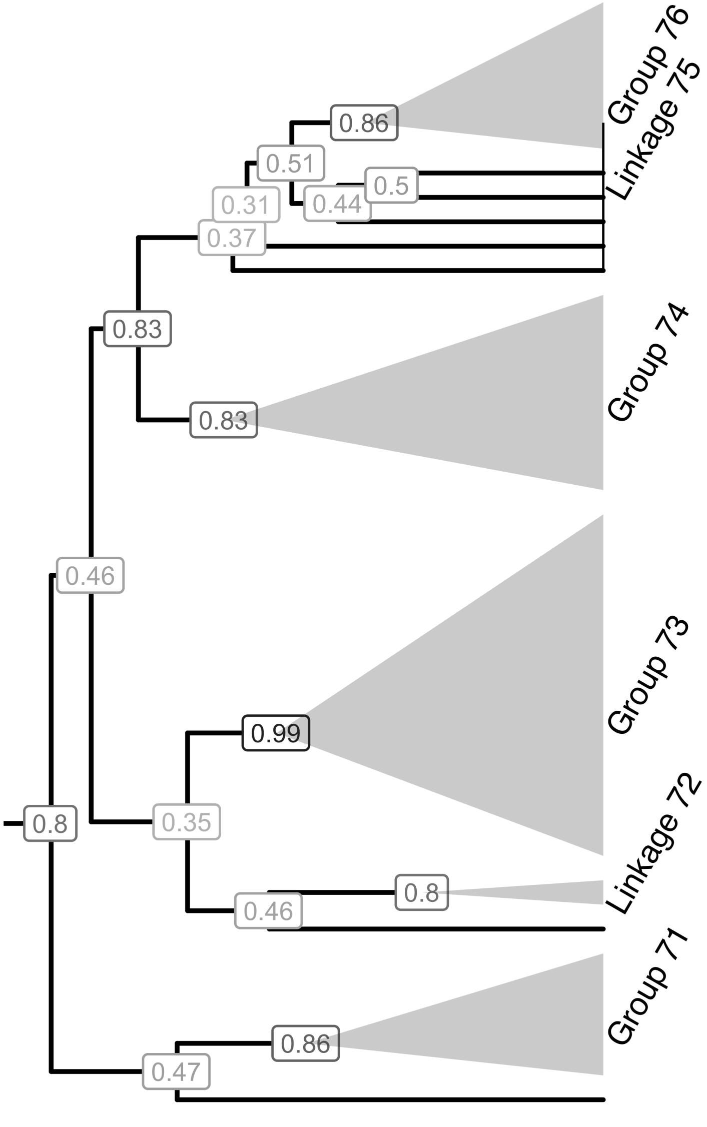


Figure 2.14: Subgroups of Mixtec Group 7 with node posteriors

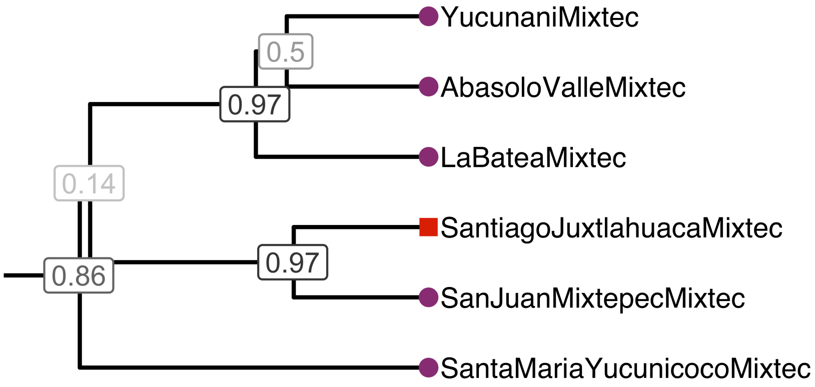


Figure 2.15: Mixtec Group 7.1 (purple = Josserand's (1983) Mixtepec, red = Josserand's (1983) Southern Baja)

level group combining those two – in other words, there is no strong evidence for a higher level Southern Baja-Guerrero group.

There are, however, well supported lower divisions. Subgroup 7.3.1 covers Josserand's (1983) Guerrero group, except for Tlahuapa and including the Southern Baja variety of Tepango. The village of Tepango is one of the westernmost Mixtec villages, surrounded by Nahuatl and Mè'phàà speakers (Josserand 1983:105). It is an important variety for the reconstruction of proto-Mixtec because it is one of the very few that preserves final glottal stops. It is classified by Josserand as belonging to the Southern Baja group based on the vowel correspondences. A detailed study of other sound changes is outside the scope of this paper, but based on the data we collected, Tepango does show a closer affinity to Yoloxóchitl Mixtec (Guerrero) than Coicoyán Mixtec (Southern Baja). We thus suggest that Tepango is in fact placed correctly in our tree. The situation is different for Tlahuapa Mixtec. Based on current knowledge, this variety should be part of the same group as the other varieties spoken in Guerrero (Subgroup 7.3.1). The reason for the 'misplacement' probably has to do with the circumstances of data collection. The data for the Tlahuapa, Piedra Azul, and San Marcos de la Flor varieties were all collected from diaspora speakers in California in a collaborative documentation project

(Campbell & Reyes Basurto forthcoming). Most likely, the speakers influenced each other in selecting certain entries over others for some of the meanings, leading to a higher number of 'apparent' cognates between those three varieties – or in other words, obfuscating semantic shifts between 'cognates'. This illustrates well the importance of a detailed knowledge of the data sources and collection techniques and the need for broad-scale, rigorous language surveys.

Within Subgroup 7.3.1, there is a well supported split that sets Alcozauca Mixtec apart from the other varieties. In Glottolog (Hammarström et al. 2021), the varitety of Xochapa is listed as a dialect of Alcozauca Mixtec, but our study suggests that this classification should be revisited. Next, we find a well supported division into two groups. The smaller group contains the varieties of Tepango and Yoloxóchitl, while the second, larger group (or 'northern' division) contains the rest of the varieties not already mentioned, to the exclusion of Xochapa which seems to occupy a position in between those divisions.

The rest of Group 7.3 covers varieties previously classified as Southern Baja. All of these varieties are spoken in the northwestern part of the state of Oaxaca very close to the border with Guerrero. These varieties are thus located in a geographically contiguous area with those of Subgroup 7.3.1, so it is not surprising that they would have a closer relationship to each other than previously assumed.

Linkage 7.2 proper contains varieties spoken further east from those of Subgroup 7.3 and are also classified as Southern Baja by Josserand (1983). However, the next node with a high posterior (0.99) does not neatly separate the varieties into these two groups, but rather into a larger group consisting of the Guerrero varieties as well as the westernmost varieties of the Southern Baja (centered around San Martín Peras and Coicoyán de las Flores). It is unclear whether the variety of Ixpantepec should be included in this group given the low posterior connecting it to the varieties of Tecomaxtlahuaca and Duraznos, but based on Josserand's classification and our own fieldwork in the area, we conclude that is does form part of this

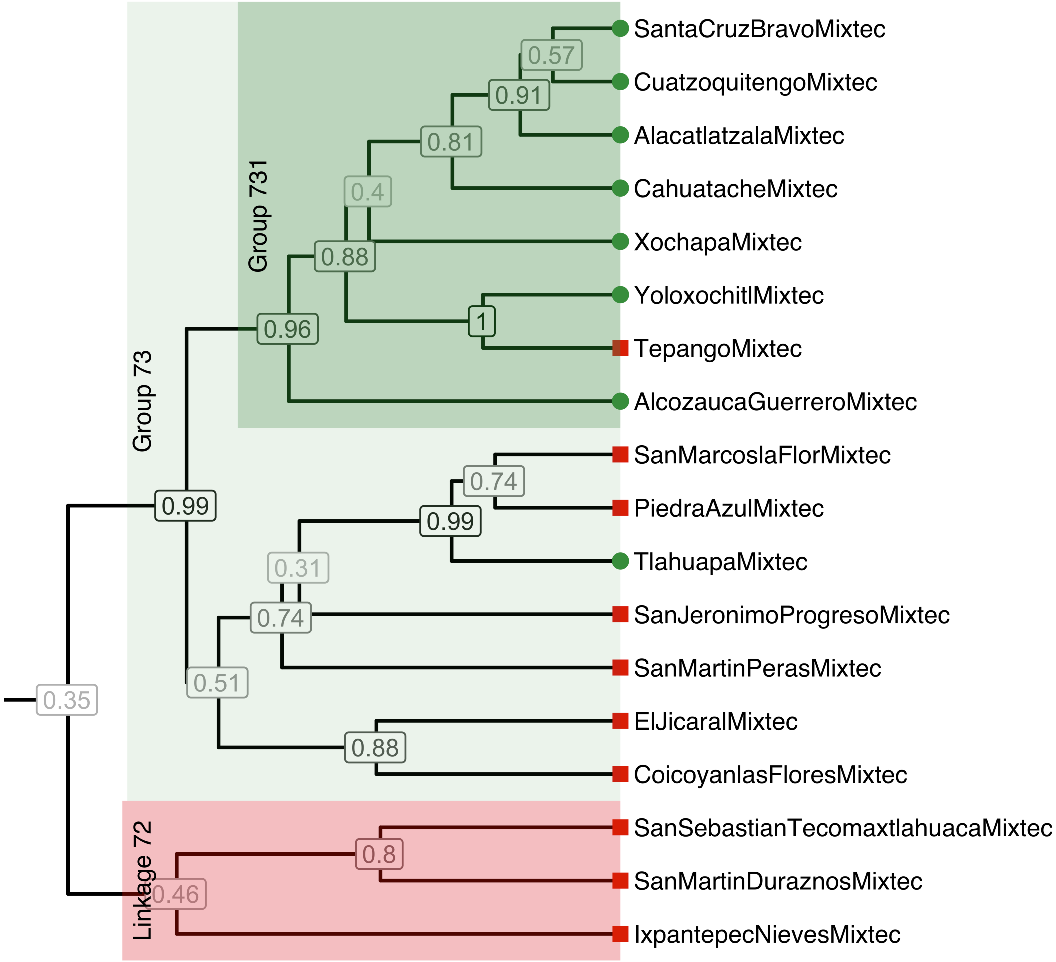


Figure 2.16: Mixtec Linkage 72 and Group 73 (green = Josserand’s (1983) Guerrero, red = Josserand’s (1983) Southern Baja)

linkage.

What remains unclear and needs further investigation is how Group 7.1, Linkage 7.2, and Group 7.3 are connected to each other. The posterior of the node linking them together is low, which means that it is questionable whether a group covering all of them should be posited. The varieties of this region, called the Mixteca Baja, are not as well studied and documented as those of the Mixteca Alta or the Coast and this was even more the case at the time of Josserand’s (1983) study. It is thus possible that more evidence for this new division will be found based on future research.

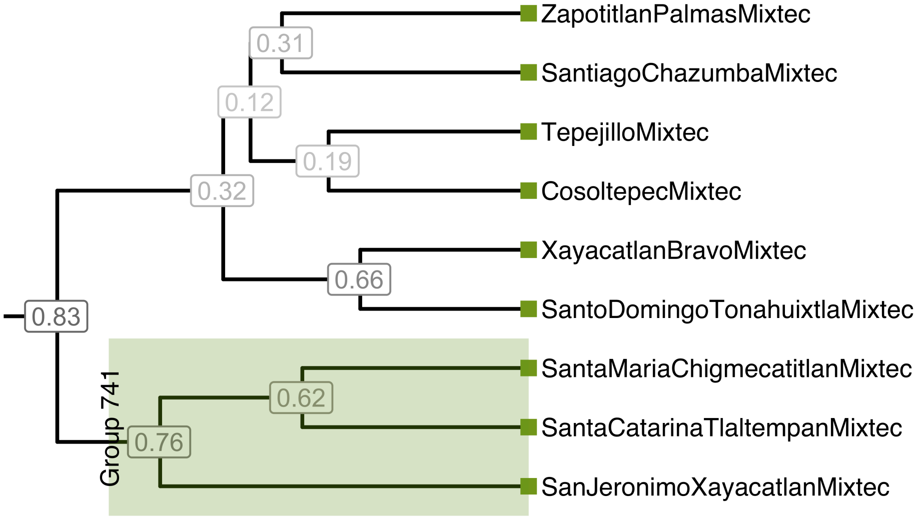


Figure 2.17: Mixtec Group 74 (Josserand's (1983) Northern Baja)

Subgroup 7.4 corresponds to the Northern Baja group from Josserand (1983). Internally, there is a division (Subgroup 7.4.1) that sets apart the varieties of Chigmecatitlan, Tlaltempan, and Jerónimo Xayacatlán from the others, cf. Figure 2.17. With respect to the first two, this corresponds well with the geographic location, being the northernmost Mixtec varieties, spoken in the state of Puebla surrounded by Nahuatl and Popoloca speakers (Josserand 1983:105). The placement of Jerónimo Xayacatlán needs to be investigated further. This variety is spoken further south near the border to Oaxaca and in very close proximity to Tonahuixtla and Xayacatlán de Bravo, with which we would expect it to be grouped together. The relationships of the rest of the varieties of Subgroup 7.4 remain unclear with very low posteriors connecting them.

Linkage 7.5 consists of a number of varieties that cannot be clearly linked together or connected to Group 7.6. The varieties included in this agglomerate are those classified as the Tezoatlán group by Josserand (1983) plus the variety of Atenango from her Central Baja group, cf. Figure 2.18. Subgroup 7.6 covers Josserand's Central Baja and Western Baja varieties. The internal structure is not very well supported, but there is a smaller group with a relatively

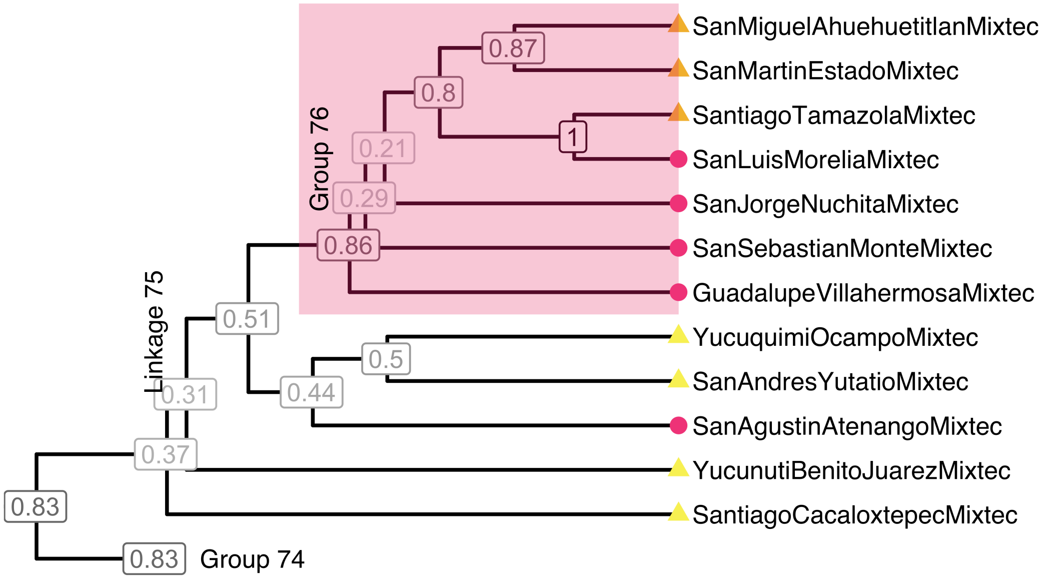


Figure 2.18: Mixtec Linkage 75 and Group 76 (yellow = Josserand’s (1983) Tezoatlán, pink = Josserand’s (1983) Central Baja, orange = Josserand’s (1983) Western Baja)

high posterior covering Josserand’s (1983) Western Baja varieties and the Central Baja variety of Morelia.

We summarize our results in Table A.2 in the Appendix and Figures 2.19 and 2.20, as a new sub-grouping proposal for Mixtec, which we see as a much needed, up to date starting point for further research on the history of these languages. The first three splits in our MCC tree identify varieties that are either temporally removed (in the case of Teposcolula) or groups that migrated to areas not originally inhabited by Mixtec speakers (Group 1 to the far north and Group 2 on the coast). This history of relative separation from other Mixtec speakers is well reflected in our results. With respect to the dialect areas proposed by Josserand (1983), few of our groups completely overlap with her areas. Our results diverge most from hers in proposing a large high-level group (Group 7) comprising seven of the dialect areas and with respect to the internal structure of the subgroups and linkages (Groups/Linkage 71-76).

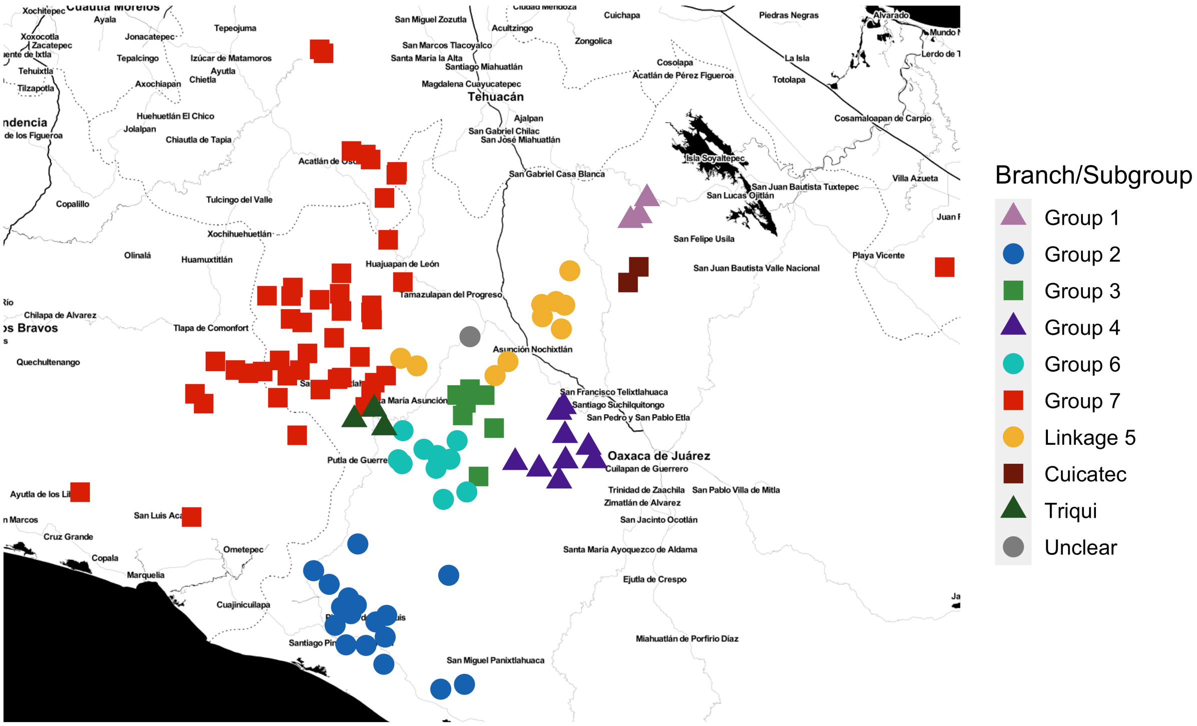


Figure 2.19: Map of the sample languages with colors/shapes indicating primary branches and new Mixtec subgroups proposed by our model

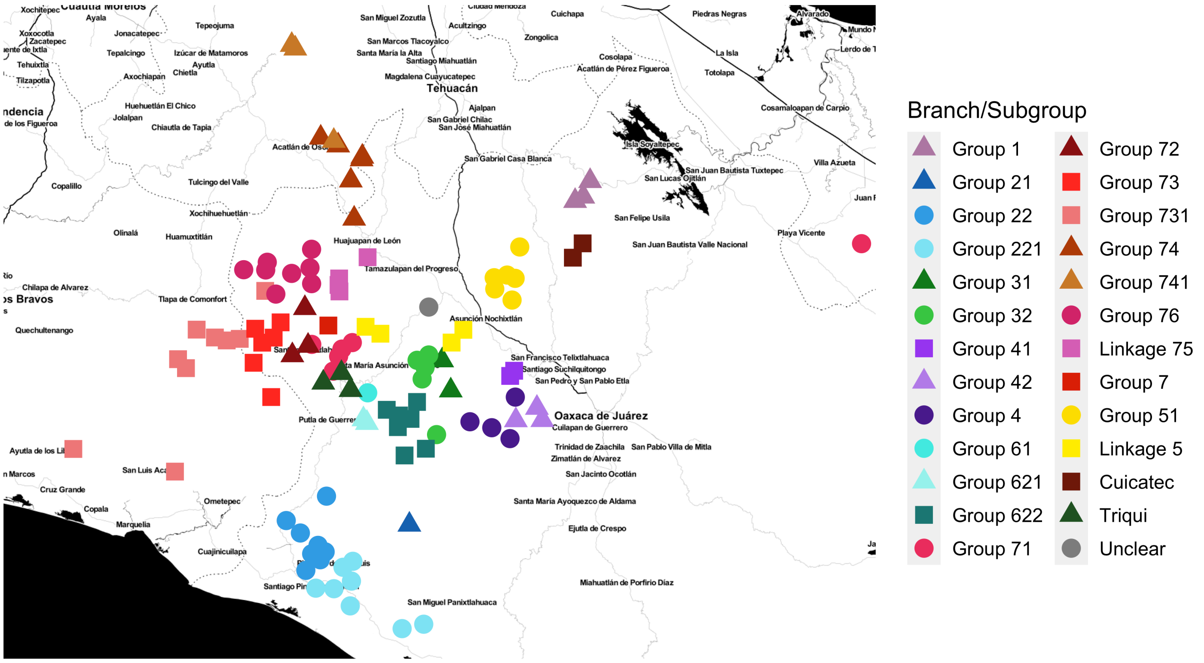


Figure 2.20: Map of the sample languages with colors/shapes indicating primary branches and new Mixtec subgroups with lower level divisions proposed by our model

2.6 Conclusion

Despite claims to the contrary, our study shows that much can be learned about Mixtecan language history by applying and combining traditional historical linguistic methods (such as establishing cognate sets) and Bayesian phylogenetics. Our results indicate that at least certain sets of varieties within this language family are best viewed as linkages or dialect chains, but this does not mean that we cannot investigate their genealogical relationships any further. We provide a starting point to further evaluate the linguistic relationships within these linkages as well as their relationship to other groups. In addition, we recover many well supported, coherent groups within the Mixtec branch, suggesting that some of its history can be described as relatively tree-like.

This adds further support to the idea that the Mixtecan language family can neither be characterized completely as a tree nor accounted for solely in a wave model. In Mixtec, we identify four groups (Group 1, 2, 4, 6) which are very well supported by the MCC tree and two more (Groups 3 and 7) which are relatively well supported. What is less clear, as mentioned in Section 2.5, is the relationship between those groups. The same is true internally of some of the larger groups, i.e. there are some well-supported lower-level subgroups, while other varieties form part of loosely connected linkages (cf. the internal structure of Group 7). This could reflect a scenario in which more wave-like periods of diversification were in turn followed by more tree-like ones.

The limitations of our study, such as the exclusion of verbs and the absence of good calibration points, also leaves some questions unanswered. One of these concerns the relationship between the three primary branches of Mixtecan. Since our model grouped Triqui and Cuicatec together – probably due to Mixtec influence on both – we cannot confirm or discard the idea that Mixtec and Cuicatec are more closely related to each other than either is to Triqui. We would also expect that some of the varieties placed outside linkages or groups can be re-evaluated for group membership once there is a better understanding of sound correspondences and sufficient analysis of verbs so that they can be integrated into the cognate database.

In Section 2.1, we mentioned a recent small-scale intelligibility study carried out around the town of Juxtlahuaca Padgett (2017) and mentioned the discrepancies to classifications in reference catalogs (Lewis et al. 2015, Hammarström et al. 2021). There is no reason to believe that the result would be much different for other parts of the Mixteca, demonstrating the need for a solid basis for language classification that involves not only traditional classification methods, but also studies and reports of intelligibility.

We also show that a good knowledge of data sources is crucial for the ability to correctly interpret the maximum clade credibility tree, especially with respect to varieties which appear in unexpected places in the tree. This in turn means that for language families with little historical data available, the best results can be achieved by gathering data through large-scale surveys applied consistently. This could help eliminate interference from differences in data gathering and preparation. The last such survey was conducted in the Mixteca region in the late 1970s (Josserand 1983:xi), focusing on Mixtec. We hope that our work will inspire a much needed update and expansion of this work.

2.7 Data Availability

The data underlying this chapter and all supplementary materials are available at `https://osf.io/n3uev/?view_only=32d258b0b80e437a85d3ab88255790be`.

Chapter 3

Patterns and distributions of sound change in Mixtec

Disclaimer

A slightly modified version of this chapter was submitted to the Journal of Historical Linguistics as: Auderset, Sandra & Eric W. Campbell. Patterns and distributions of sound change in Mixtec

Author contributions: SA: Conceptualization, Data curation, Formal analysis, Methodology, Writing - original draft, Writing - review and editing; EWC: Writing - original draft, Writing - review and editing.

3.1 Introduction

Understanding the history of a language family improves our accounts of the synchronic variation we observe, of the relationships between languages, and our general knowledge of processes of language change. But the importance of unraveling the diachrony of a language family goes far beyond that. It provides a window to the past of these people, often adding detail that is otherwise inaccessible. This is particularly true in areas like Mesoamerica, where we find complex patterns of interaction between many ethnolinguistic groups that still hold many open questions for population genetics and archaeology. Historical documents of languages are sparse, even though Mesoamerica was one of few sites of independent innovation of writing. The comparative method of historical linguistics based on the principle of regular sound change is still the primary tool for establishing language families, 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 (cf. Bowern 2018, Greenhill et al. 2020, List et al. 2018, among others). It is sometimes assumed that these methods, which result in family trees, do not apply well in situations of continued contact between languages, that is, in so-called dialect areas or dialect continua (Ross & Durie 1996, Kalyan & François 2018).

The Mixtec language family has been characterized as a dialect continuum, (Longacre 1957, Josserand 1983), and previous studies show that reconstruction and identification of sound changes is possible in this language group, just as in any other (Rensch 1976, Longacre 1957, Bradley & Josserand 1982, Josserand 1983, Dürr 1987, Kaufman in press). Nevertheless, the impression remains that because we are dealing with a dialect continuum, we cannot go beyond positing dialect areas (Josserand 1983). The problem with dialect areas based on overlapping isoglosses is that they are based on selected sound changes deemed important and as such necessarily result in an incomplete account of language relationships and histories. 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. This is true of both their linguistic history, including complex relationships between varieties and unclear higher-level groupings, and – connected to that – their migration history, which often includes whole towns relocating to another area (Chance 1986).

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 is most closely related to the Cuicatec and Triqui languages, and these three groups together comprise the 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 (Rensch 1976, Kaufman 2006, Campbell 2017b). We present an updated analysis of Proto-Mixtec lexical reconstructions and sound changes, focusing on the distribution of those changes in space and time and what this tells us about the history of the Mixtec languages. There is now an extensive database of Mixtec cognate sets available, as well as an up-to-date phylogeny of the language family (Auderset et al. submitted). This means that we are in an ideal position to re-evaluate earlier reconstructions and sound changes. We also assess how well the sound changes we identify align with Josserand's (1983) dialect areas and the subgroups proposed by a Bayesian phylogenetic study (Auderset et al. submitted). The Mixtec language family consists of perhaps some 200 distinct local varieties[1] that form a continuum, many of which are not mutually intelligible; Smith Stark (1994) counts 189 distinct municipalities in which Mixtec is spoken. This large number of varieties poses a challenge 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).

Our work builds on previous Mixtec(an) reconstructions and classifications obtained with the comparative method and quantitative methods. Longacre's (1957) Proto-Mixtecan reconstruction, incorporated in Rensch's (1976) comparative Otomanguean, was based on only one

[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).

variety each of Cuicatec and Triqui and just four Mixtec varieties. He reconstructed final (originally prominent) syllables for Proto-Mixtecan. Mak & Longacre (1960) updated this reconstruction with an expanded sample of 28 Mixtec varieties. Bradley & 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 & Longacre's (1960) in important ways (see Campbell 2017a:10). Josserand's (1983) subsequent work remains the state of the art in Mixtec reconstruction. Focusing on vowel correspondences and changes, she compared data from 120 Mixtec varieties, reconstructed 188 Proto-Mixtec forms, and classified Mixtec languages into twelve 'dialect areas', for some of which she suggests possible lower-level divisions. Kaufman (in press) revisits and revises 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.

Since Josserand's (1983) seminal work, several additional Mixtec varieties have been documented and described. Moreover, our approach includes computational and quantitative methods that were not available at the time of the previous studies. With these advances in data and methods, we are able to incorporate a much larger and more representative and updated data set to inform Proto-Mixtec reconstruction and sound changes, which are of interest to scholars of Otomanguean and Mesoamerican languages, as well as to communities who are interested in developing and sharing materials for language pedagogy and maintenance. Our segmental reconstruction largely accords with Josserand's (1983), and therefore Kaufman's (in press), requiring no substantial revision to the Proto-Mixtec sound system.

Our goal is not to explain each reflex in each variety in detail, but to identify broader patterns in the sound changes, resolve selected outstanding questions, point to areas in need of further work, and finally, to consider the results in light of previous classifications: Josserand's

(1983) 12 Mixtec dialect areas (reproduced in Figure 3.1), and Auderset et al.'s (submitted) recent maximum clade credibility (MCC) tree arrived at by applying Bayesian phylogenetic methods and the comparative method to cognacy data (reproduced in Figure 3.2). The subgrouping based on the MCC tree from Auderset et al. (submitted) is provided in the supplementary materials, since the graphic is too large to be legibly displayed on a page. We do not discuss these groupings or their relationship to Josserand's dialect areas in detail since this information can be found in Auderset et al. (submitted). As opposed to earlier studies (Longacre 1957, Bradley & Josserand 1982, Josserand 1983), we do not pre-select varieties or sound changes deemed representative, but rather work with the full set of Mixtec data available to us. Given the data, we establish cognate sets and reconstruct corresponding proto-forms. We then analyze the correspondences across the sample languages and establish fine-grained sound changes to account for the modern reflexes. We do this by combining methodology from historical linguistics and typology, creating reproducible and re-usable interlinked databases. Due to limitations of space, we cannot discuss all developments of each Proto-Mixtec sound in detail in the paper, but the complete data for doing so is provided as supplementary material. Given the large number of languages and changes analyzed it is challenging to adequately summarize the data and identify tendencies. To this end, we apply various data visualization and aggregation methods and measures, which allow us to evaluate the information gathered here in terms of earlier studies. More specifically, we explore how well the distributions of sound changes align with previous classifications of Mixtec languages. This is of particular interest because of the view of the Mixtec language family as a dialect continuum. Dialect continua are characterized by changes which do not neatly separate groups of languages from each other, but rather arise in a language or languages – the center of innovation – and spread out from there. The center of innovation can shift depending on the change, as does the extent to which the change 'travels' from the center. If sound changes in Mixtec predominantly proceed in this way, we would expect them to generally align with the previously established dialect areas (Josserand 1983), but not as much with the subgrouping classification based on a tree (Auderset et al. submitted). As we show below, this is true for some changes, but not for the majority.

3.2 Data collection and reconstruction

The data that serves as the basis for our study constitutes a subset of a previously established data set (Anonymous 2022) focusing on subgrouping within the Mixtecan language family as a whole, from which we extracted all the Mixtec language data. The data was collected with a list of 209 concepts of basic vocabulary. We exclude verbal forms because they require aspect-mood inflection, which is not well understood on a comparative Mixtec level and is not always provided or reliably identifiable in the sources. Details about the construction of the original word list can be found in Auderset et al. (submitted), and the complete list is also reproduced as supplementary materials X. Our sample includes 104 Mixtec languages; an overview of all languages and sources can be found in the supplementary materials and in Figure 3.2, which shows their location as well as subgroup membership according to the previous study (Auderset et al. submitted). The entries were originally collected in the orthography provided in the source and then converted to a standardized representation in IPA using orthography profiles (one for each source), also standardizing tone notation. The details of the conversion and standardization are laid out in detail in SM 2 of Auderset et al. (submitted).

Cognate coding was carried out based on the comparative method informed by previous reconstructions. 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). As mentioned in Section 4.1, the phoneme inventory for Proto-Mixtec previously proposed by Josserand (1983), among others, is maintained and is summarized in Table

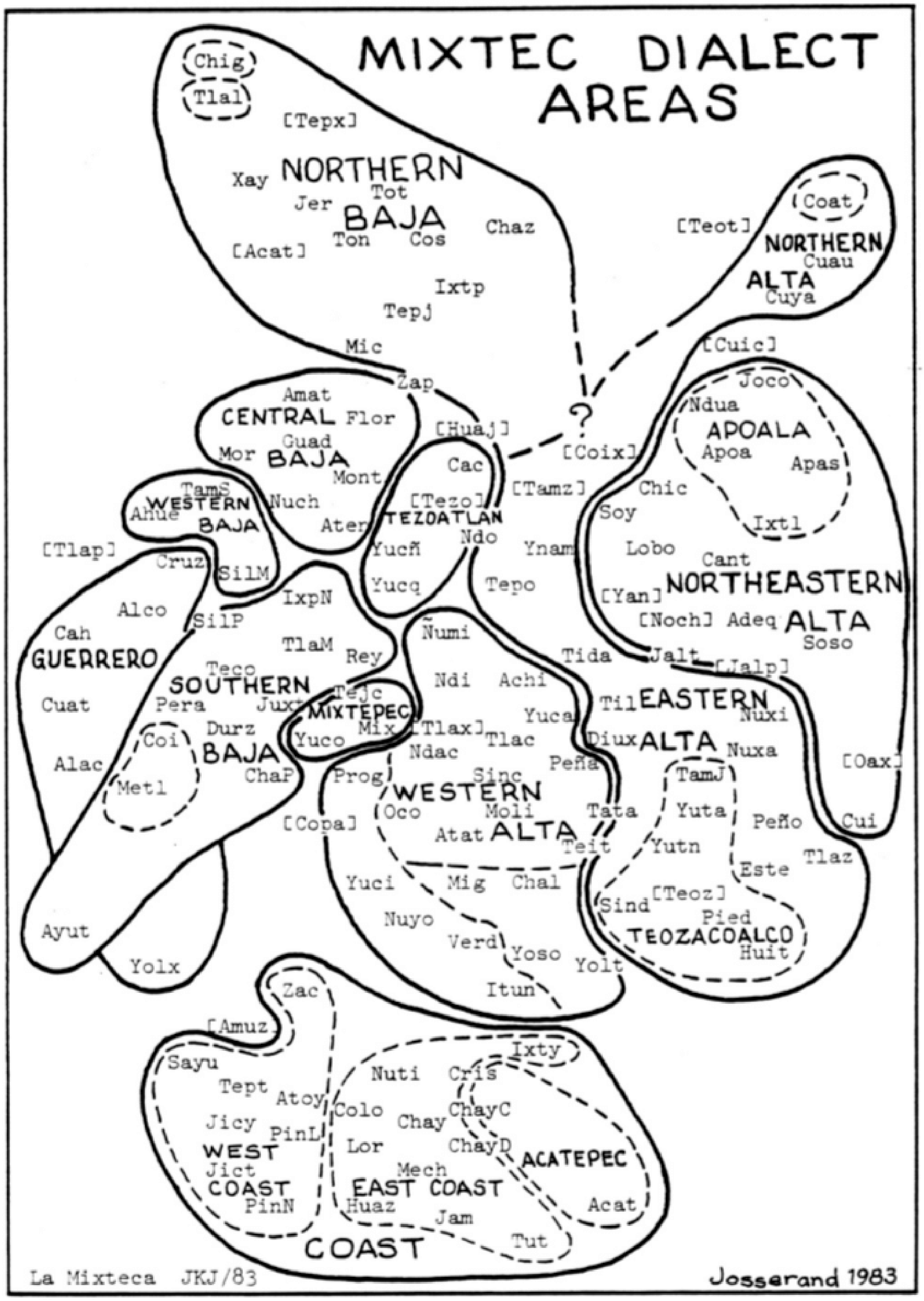


Figure 3.1: Dialect areas with subdivisions by Josserand (1983:470)

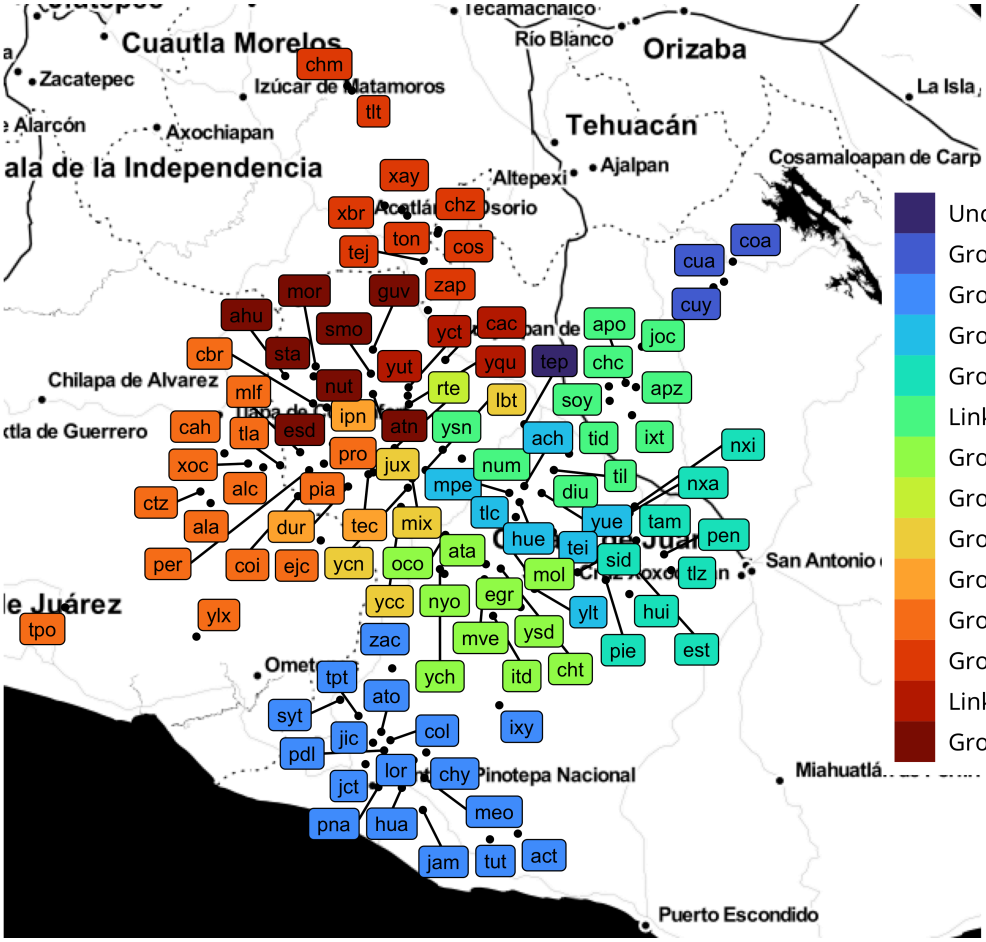


Figure 3.2: Map overview of sampled varieties colored by subgroup (Auderset et al. submitted)

Table 3.1: Proto-Mixtec phoneme inventory

	dental	velar	lab.-vel.	glottal		front	central	back
plosive	*t*	*k*	k^w	ʔ	close	*i*	*ɨ*	*u*
prenas. pl.	^{n}*d*				mid	*e*		*o*
nasal	*n*				open		*a*	
fricative	*s*	*x*						
approximant	(*l*)	*j*	*w*		suprasegmentals:		nasalization; tone	

3.1. Our Proto-Mixtec segmental inventory differs from Josserand's (1983), but agrees with Kaufman's (in press), in that we include the glottal stop as a consonant, while Josserand (1983) considers laryngealization to be a vocalic feature.[2] However, the diachronic behavior of the glottal stop is such that our results would be no different if we analyzed it as a vowel feature.

We bracket Proto-Mixtec *l because it is reconstructed by Josserand (1983) and others, but almost all the forms reconstructed with this proto-phoneme have modern reflexes which alternate between *l* and *s* or more rarely between *l* and ^{n}*d*. The alternation is not predictable and does not apply consistently across the lexicon of any one variety. It thus cannot be characterized as a sound change, which is why we will not discuss it any further in this paper. According to Kaufman (in press), Amuzgo, the language group most closely related to Mixtecan, displays 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 residue of these prefixes. In some modern Mixtec varieties, *l* has acquired a status of a phonaestheme for small or cute things or animals (Mendoza Ruíz 2016). The forms reconstructed with *l and a summary of the modern reflexes are provided in Table 3.2. Cognate set 674 SMALL (SINGULAR) and 91

[2] The synchronic phonological representation of laryngealization in Mixtec languages has received considerable attention and divergent analyses. For example, Castillo García (2007), McKendry (2013) and Hinton et al. (1991) also treat laryngealization as a vocalic feature, in the Mixtec varieties of Yoloxóchitl, Nochixtlán, and Chalcatongo, respectively, as does Gerfen (1996) for Coatzospam Mixtec, where it is automatically inserted word-medially. Macaulay & 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 and Alcozauca Mixtec, respectively. North & Shields (1977) and Pike & Cowan (1967) consider it to be a glottal stop consonant in Silacayoapan and Huajuapan Mixtec, respectively.

Table 3.2: Forms reconstructed with *_l_ by Josserand (1983)

ID	Concept	PMx	l-reflexes	Other reflexes	
91	CAT	*_wiluʔ_	48	-	
674	SMALL (SINGULAR)	*_luʔu_	28	-	
49	BIRD	*_laa_	24	*s and its reflexes	52
263	FROG	*_laʔwa_	21	*s and its reflexes	47
574	PUS	*_lak^{w}aʔ_	18	*^{n}d and its reflexes	83
826	URINE	*_lele_	11	*s and its reflexes	10
660	SMOOTH	*_lɨʔwɨʔ_	9	*^{n}d and its reflexes	64
194	EAR	*_loʔo_	4	*s and its reflexes	93
635	TAIL	*_luʔwẽʔ_	2	*s and its reflexes	65

CAT are the only ones 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 *^{n}_d_. We therefore do not reconstruct *_l_ in the proto-forms except for SMALL. We actually identified other cognate sets for SMALL (both for plural and singular) that show consistent reflexes for *_l_, namely 668 SMALL (PLURAL), which we reconstruct as *_waliʔ_, and 673 SMALL (SINGULAR), we reconstruct as *_luʔ^{n}di_ or alternatively *_luti_.[3]

For each cognate set with sufficient data, we reconstructed a Proto-Mixtec form. For many of the sets, previous reconstructions are available from Josserand (1983) and Dürr (1987). We re-evaluated their reconstructions in light of the data available to us and for the most part no adjustment was necessary. For a few forms, we propose adjustments, mostly with respect to the vowels *_e_ and *_u_, which are difficult to distinguish from *_a_ and *_o_, respectively. Below, we briefly summarize the rationale for each of the changes to the proto-forms.

In set 476 NEW, we added a final glottal stop: *_xeeʔ_. Dürr (1987) lacked the relevant data from Tepango, which retains the final glottal stop, and the improved proto-form has no impact on the assessment of segmental changes in this word.

In three sets, we reconstruct *_õ_ where earlier sources reconstructed *_ũ_. The affected sets

[3] The cognate set for the latter has few entries that show a lot of variation with respect to the medial consonant, hence the two possible reconstructions.

Table 3.3: Example reflexes of cognate sets with **eje* and **eji*

Variety	SLOW (662)	HEAVY (326)	MAN (414)
Proto-Mixtec	**kʷ e j i*	**w e j i*	**t e j e*
IxpantepecNieves	*kʷ e e*	*β e e*	*tʃ a a*
YucuquimiOcampo	*kʷ e e*	*β e e*	*t a a*
SantoDomingoHuendio	*kʷ e e*	*β e e*	*t e e*

are 93 BLACK **toõʔ*, 79 BOX **xetõʔ*, and 369 IRON **toni*. The last is likely a borrowing from archaic Spanish *tomín* 'silver coin' used in parts of the Americas and ultimately a measure term of Arabic origin. The difficulty in reconstructing **õ* versus **ũ* stems at least partially from the fact that in many varieties, the reflexes of the former merge completely with the latter to *ũ* (or *u* with loss of nasalization). There is also one variety, San Andres Yutatio, that went in the opposite direction, merging **õ* and **ũ* as *õ*. Neither of the two back nasal vowels can be posited for all of the relevant reconstructions and account for all of the reflexes, as about a quarter of the varieties sampled display mixed reflexes. Unfortunately, the patterns are varied and difficult to disentangle at this stage; they require further research, including consideration of tone as a conditioning factor.

In set 662 SLOW, we reconstruct **kʷeji* instead of **kʷeje* as in earlier sources, because the reflexes in our data set fit better with those of set 326 HEAVY **weji* than with those of 414 MAN **teje* in varieties that do not retain **e* in that context, illustrated in Table 3.3.

Similarly, for set 103 CHILD, Josserand (1983) reconstructed **saʔji*, while we reconstruct **saʔje*. The reason is that the reflexes of the final syllable do not fit those of other Proto-Mixtec forms with the same shape in some varieties, illustrated in Table 3.4. It is possible to attribute these differences – the retention of *j* and/or the differing final vowel – solely to the vowel in the first syllable, in keeping with Josserand's reconstruction. However, the reflexes of varieties like Coatzospam and Yucuhiti Mixtec, where the initial vowel is identical across all three sets, are easier to explain with our modified reconstruction.

Table 3.4: Example reflexes of cognate sets with final *ʔje and *ʔji

Variety	CHILD (103)	DISEASE (172)	HOUSE (363)
Proto-Mixtec	**s a ʔ j e*	**k*^{w} *e ʔ j i*	**w e ʔ j i*
SantaCatarinaEstetlaMixtec	*ð a ʔ ʒ a*	*k*^{w} *e ʔ e*	*w e ʔ e*
SantaMariaYucuhitiMixtec	*s e* ^{1} *ʔ j a* ^{3}	*k*^{w} *e* ^{1} *ʔ i* ^{1}	*β e* ^{3} *ʔ i* ^{3}
SanJuanCoatzospamMixtec	*i* ^{5} *ʔ ʃ a* ^{5}	*k*^{w} *i* ^{5} *ʔ i* ^{5}	*β i* ^{1} *ʔ i* ^{1}

We have also re-evaluated two sets with respect to the labio-velar **k*^{w}. For set 136 COMB, earlier sources reconstructed **kuka*, a form that is attested as such in a number of languages. The rest of the reflexes show *k*^{w}*ika*, which is the form that we reconstruct. The main reason is that it is easier to explain the change from **k*^{w}*i* to *ku* than the reverse given other changes that need to be posited independently of this cognate set. In many varieties, **wi* in final syllables develops into a reflex *u* (see Section 3.5.1) and the change from **k*^{w}*i* to *ku* can be seen as parallel to that. Moreover, it is more difficult to explain a change from **u* to *i* in the absence of a suitable environment for vowel fronting, such as adjacency to the palatal semivowel. Conversely, for set 159 CROSS-SEX SIBLING, the previous reconstruction is **k*^{w}*aʔa*, while we reconstruct **kuʔwa*. Both of these forms are attested in about an equal number of modern languages. The reason for changing the reconstruction in this case is again that given other sound changes operating in these languages, it is easier to explain the change from **ku* to *k*^{w}*a* in this context, summarized in Table 3.5. Both the *o* and *i* reflexes in the first syllable are congruent with other changes involving **u* (cf. Section 3.5.3). Furthermore, the loss of **w* after a glottal stop is well attested, while fission of *k*^{w} into a plain stop and movement of a resulting semivowel from the first to the second syllable is not.

Perhaps the most drastic change we propose is in cognate set 260 FOUR. Josserand (1983), and Dürr (1987) following her, reconstructed **kɨwɨ̃ʔ*. Their reconstruction is based on seven varieties that do show reflexes with these vowels, see upper half of Table 3.6. However, the vast majority of reflexes look quite different with respect to the first vowel. A representative

Table 3.5: Example reflexes of 159 CROSS-SEX SIBLING

Variety	CROSS-SEX SIBLING (159)
Proto-Mixtec	**k u ʔ w a*
MetlatonocMixtec	*k u ʔ β a*
SanMartinDuraznosMixtec	*k o ʔ β a*
AlacatlatzalaMixtec	*k i ʔ β a*
SanJeronimoXayacatlanMixtec	*k u ʔ a*
SanAndresYutatioMixtec	*k i ʔ o*
SanPedroYosonamaMixtec	k^{w} *a ʔ a*

sample of those reflexes are given in the lower half of Table 3.6, where the first vowel is a back vowel. Furthermore, the interpretation of vowel qualities in the colonial era documentation of the Teposcolula variety is difficult, and cannot be rechecked with speakers, so reconstructions should not be based primarily on this variety. Of the other six varieties that exhibit *i* in the first syllable, Coatzospam, Cuauhtemoc, and Cuyamecalco form a close-knit genetic unit (Group 1), and we are likely observing evidence of a single shared innovation. The remaining three varieties all belong to Group 4. Both of these groups exhibit other isolated vowel changes to *i* (cf. Sections 3.5.6 and 3.5.5). We thus reconstruct **kowĩʔ*, which accounts much better for the majority of reflexes, while the 'exceptional' reflex in Group 1 and the varieties of Group 4 can be accounted for by positing changes similar to others found in these varieties, or to vowel harmony, which is common throughout Mixtec. Moreover, the reflexes with non-back vowels do not fit with cognate sets 163 DAY and 204 EGG, which would have the same environment except that the final vowel is not nasal, a context not elsewhere seen to condition vowel backing in Mixtec. Table 3.6 demonstrates that for the majority of varieties, (in the lower half), DAY and EGG display identical vowels, while the vowels in FOUR are different.

Table 3.6: Example reflexes of cognate sets FOUR, DAY, and EGG

Variety	FOUR (260)	DAY (163)	EGG (204)
Proto-Mixtec	**k o w ĩ ʔ*	**k ɨ w ɨ ʔ*	*^{n}*d ɨ w ɨ ʔ*
SanPedroySanPabloTeposcolula1600	*k e m i*	*k e β i*	^{n}*d e w i*
SanJuanCoatzospam	*k ɨ m i*	-	^{n}*d ɨ β i*
CuyamecalcoVillaZaragoza	*k ɨ m i*	-	^{n}*d ɨ w ɨ*
SantaAnaCuauhtemoc	*k ɨ m i*	-	^{n}*d ɨ w ɨ*
SanAntonioHuitepec	*x ɨ m ɨ*	-	^{n}*d ɨ w ɨ*
SanJuanTamazola	*x ɨ m ɨ*	-	^{n}*d ɨ w ɨ*
SanMiguelPiedras	*x ɨ m ɨ*	-	^{n}*d ɨ w ɨ*
ElJicaral	*k o m i*	*k i i*	^{n}*d i β i*
Tepango	*k u m i ʔ*	*k i ʔ β i ʔ*	^{n}*d i β i ʔ*
SantiagoTilantongo	*k õ õ*	*k ɨ u*	^{n}*d ɨ u*
SantaCruzItundujia	*k ũ ũ*	*k ɨ w ɨ*	^{n}*d ɨ w ɨ*

3.3 Identification and coding of sound changes

After revising or 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. The AUTOTYP project was developed to address problems that arose 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, that often fail to adequately capture a phenomenon across a large and diverse sample of languages. 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 re-usable databases. One of these principles is the use of *auto-*

typology (Bickel & Nichols 2002, Bickel 2010, Bickel et al. 2011). Autotypology 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 from 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 several 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 104 languages is comparable to that of mid-sized typological studies (Bakker 2010). Second, we work with a large amount of data points, rather than cherry-picking 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 paragraphs, we describe database creation and sound change coding in more detail.

Information about the languages of the sample is provided in the metadata file. 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, latitude and longitude of the village, subgroup membership according to two previous studies (Auderset et al. submitted, Josserand 1983), ISO-639-3-codes and Glottocodes (where applicable), our own language code, and the source(s) of the data. The full bibliographic information of the sources is provided in a bib-file.

Table 3.7: Excerpt from the proto-forms database[4]

COGID	CONCEPT	Proto-Mixtec	PMAlt.	Sources	IDJoss	IDDurr
294	GRASS	**ite*				
662	SLOW	**k^{w}eji*		k^{w}eje	163	
141	CORN	**noniʔ*		noniʔ	85	73
51	BITTER	**owe*	owa	owe	29	
6	AUNT	**sisi*		sisi		35

Table 3.8: Excerpt from the cognate database

ID	CONCEPT	DOCULECT	TOKENS	COGID
2104	CHILI PEPPER	SanJuanDiuxiMixtec	*ʒ a* ^{5} *ʔ a* ^{5}	107
19359	CHILI PEPPER	SanMartinDuraznosMixtec	*ʒ a* ^{3} *ʔ a* ^{1}	107
6524	CHILI PEPPER	SantaMariaZacatepecMixtec	*j a* ^{3} *ʔ a* ^{3} *ʔ*	107
2127	RIVER	SanJuanDiuxiMixtec	*ʒ u* ^{5} *t e* ^{5}	607
19559	RIVER	SanMartinDuraznosMixtec	*ʒ i* ^{1} *tɕ a* ^{3}	607
8487	RIVER	SantaMariaAcatepecMixtec	*j u t*^{j} *a*	607
12655	RIVER	SantoDomingoTonahuixtlaMixtec	*j o t e*	607

All reconstructed forms were entered in a database containing the cognate set ID, the reconstructed meaning of the cognate set, our reconstructed form in IPA (and where applicable an alternative possible reconstruction), the reconstructed form of earlier sources (also standardized to IPA for better comparability) and the cognate set ID (if applicable) of the earlier source(s). This database is linked with the cognate sets via the cognate set ID. An excerpt is provided in Table 3.7.

The cognate set database contains a unique identifier for each form, the language identifier, cognate set ID and meaning, the full form in IPA, as well as just the segmental and tonal reflexes (if available) separately. 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 3.8.

The sound changes were collected as a definition file containing a sound change ID, the

[4] PMAlt. = alternative Proto-Mixtec form, Sources = reconstructed form given in earlier studies, Joss = Josserand 1983, Durr = Dürr 1987

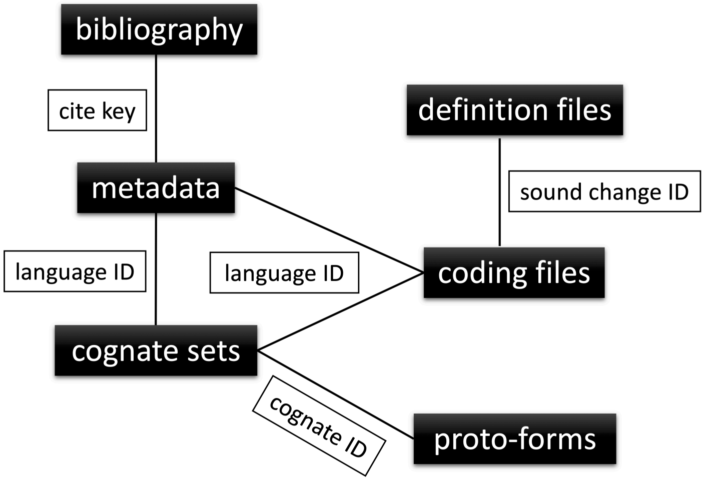


Figure 3.3: Schematic overview of the linked databases. Black boxes with white text represent databases, white boxes with black text show the unique identifiers linking two databases.

Proto-Mixtec source phoneme, the modern reflex (with phonetic variants where applicable), the preceding and following sounds or environments that may have conditioned the change, and the IDs of changes that happened before or after the change in question. The presence or absence of each sound change in each variety was recorded in the coding file. The presence of changes across varieties with partially overlapping targets and/or conditioning environments necessitated the coding of changes at a very fine-grained level. If the data were lacking or inconclusive for a particular change in a language, this is indicated with NA (not applicable). This file contains the language identifier, sound change identifier and value.

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 thus did not code for fine phonetic variants, analogical processes, and remnants of fossilized morphology. To allow the database to be expanded upon with more data in future work (for example by integrating verb paradigms), we largely refrained from specifying environments with sound classes and rather listed each conditioning phoneme separately. This no doubt has led to under-generalization for some changes, but such generalizations can be

Table 3.9: Example cognate set and derived sound change variables and coding

	TREE (799)	J13	U08	T24	U34	GS01
Proto-Mixtec	*j u t ũ ʔ	*j >ø /#_u	*u >i /_tũ	*t >n /_Ṽ	*ũ >õ /t_	*ʔ >ø /_#
S.MartinPeras	*i t ũ*	yes	yes	no	no	yes
S.MateoSindihui	*ʒ ũ n õ*	no	no	yes	yes	yes
S.MariaZacatepec	*j u t ũ ʔ*	no	no	no	no	no
St.Cacaloxtepec	*ʒ i t õ*	no	yes	no	yes	yes

recovered through later aggregation (see supplementary material). 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. A small example data set with changes derived from it is given in Table 3.9. Figure 3.3 provides an overview of all the databases and their links to each other. The files are all provided in the supplementary materials.

3.4 Consonant changes

3.4.1 Developments of Proto-Mixtec *t

Proto-Mixtec **t* in general does not exhibit many changes, and it is not regularly lost in any variety. The majority of changes concern various palatalizations before front vowels, most commonly before *i*, but also before *ɨ* (after it becomes fronted) and *e*, and in some cases before *a* or *u*. The two most common reflexes in these contexts are *tʃ* and t^j, but the details as well as the chronology of these changes are complex. We summarize the reflexes in Figure 3.4. A map with the conditioning environments is provided in the supplementary materials.

The palatalization to *tʃ* before *i* is found in all varieties of Groups 3, 4, 5, 6, and 7, but only partially in Group 1, where two varieties show *ts* as a reflex instead, and partially in Group 2. Group 2 shows a split into varieties that do not have palatalization at all, located in the west, and varieties that palatalize to t^j in various environments. Outliers are Sayultepec at the western end, which has palatalization before *i* even though all the neighboring and most closely related varieties do not, and Acatepec in the very east as the only variety in this group with a *tʃ*-reflex.

More localized developments include the change to *s* before *ɨ*, which is limited to a set of varieties of Group 74, in which the other varieties show *ts* in the same context. The varieties in this group that show the former change are geographically close to each other, which could suggest that the loss of the plosive element of the affricate is a later development from reflex *ts*. Three-way contrasts where we find a different reflex before *i*, *ɨ*, and *e* are rare and only found in two closely related varieties of Group 73 (Piedra Azul and San Marcos la Flor) and in one variety of Group 74 (Chigmecatitlan). Many of the surrounding varieties in both cases show a two-way contrast and no palatalization before *e*, so this too could be a later, localized development.

Not shown on the maps is a rarely found palatalization to t^j before *u*, which must have happened after any vowel changes to *u* to arrive at the correct reflexes. This change is only found in two closely related varieties of Group 4 (Nuxiño and Nuxaa) and in two scattered varieties of Group 7, Ixpantepec (72) and Zapotitlan (74). The latter two are most probably independendent from each other and from the change in Group 4.

Proto-Mixtec $^*t$ also undergoes changes before nasal vowels in some varieties. Most frequently it changes to t^n, but in a few varieties it turns into the nasal *n*. These changes affect a set of geographically relatively contiguous varieties around the Nochixtlan area in the eastern part of the Mixteca, but also scattered varieties in Groups 7 and one variety of Group 2. Group 3 is the only one that shows nasalization of $^*t$ throughout, which could indicate that this change originated there and subsequently spread to nearby varieties of Groups 4, 5, and 6. We currently have no good explanation for the scattered reflexes found in Groups 2 and 7, except that they must be independent innovations, which perhaps only happened quite

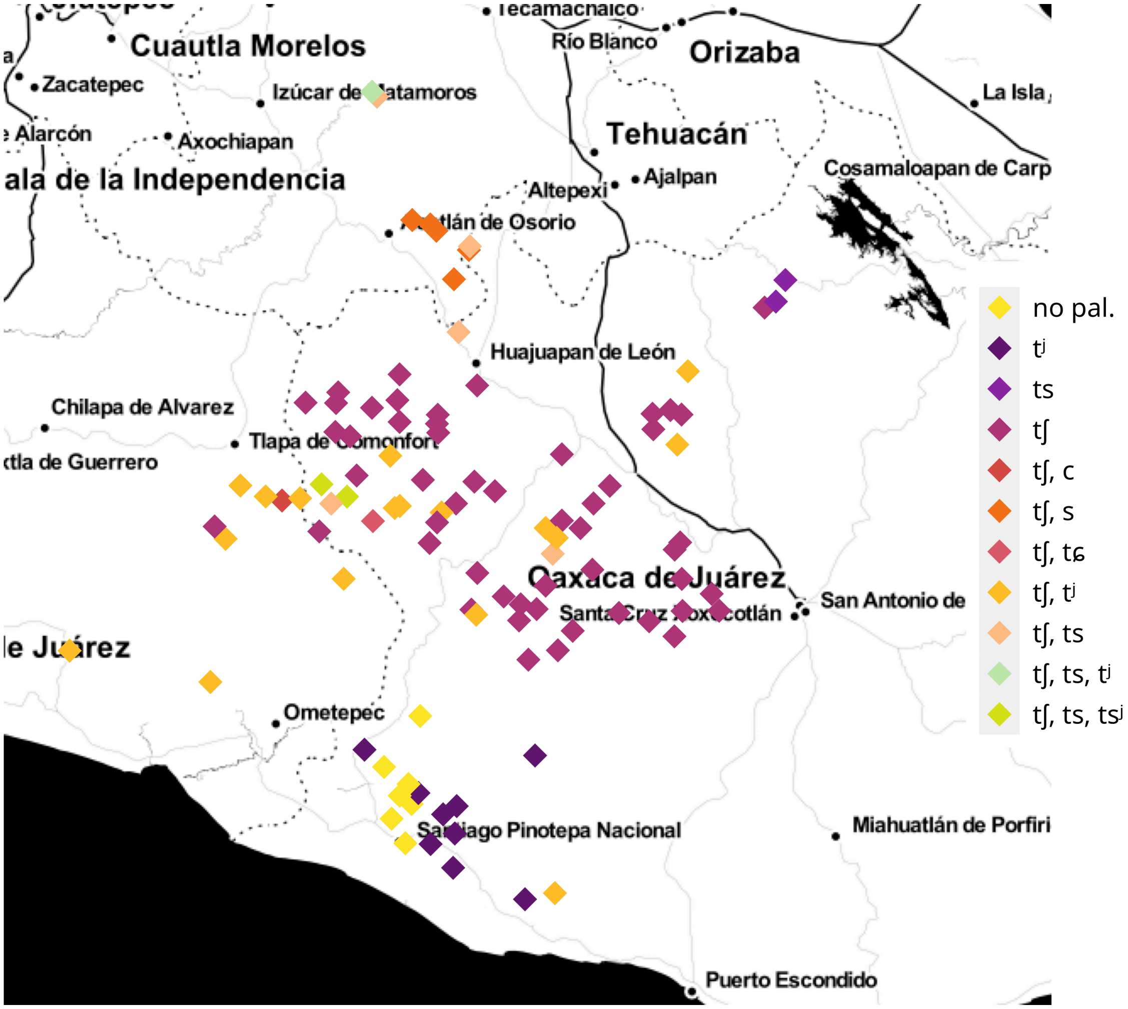


Figure 3.4: Reflexes of palatalization of *t

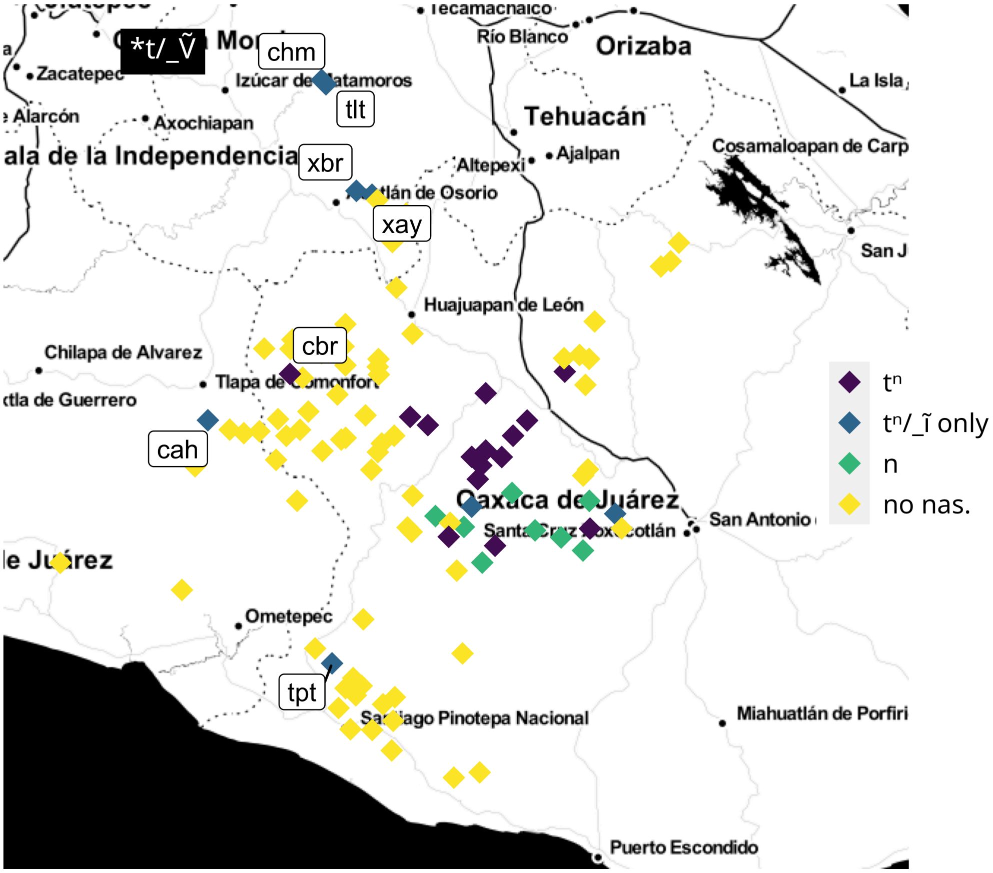


Figure 3.5: Developments of *t before nasal vowel

recently.

3.4.2 Developments of Proto-Mixtec *k and *k^w

We discuss Proto-Mixtec *k and *k^w together since they both exhibit few changes and we adjusted some earlier reconstructions with respect to these two sounds, cf. Section 3.2.

Lenition of Proto-Mixtec *k to x and *k^w to x^w occurred in a handful of languages, and this is also the case for the loss of *k before nasal vowels. Figure 3.6 shows the distribution of these changes in the varieties of Cuyamecalco from Group 1, Santiago Tamazola from Group 7, Diuxi and Tilantongo from Linkage 5, and Tamazola, Piedras, Nuxiño, Sindihui, and Huitepec

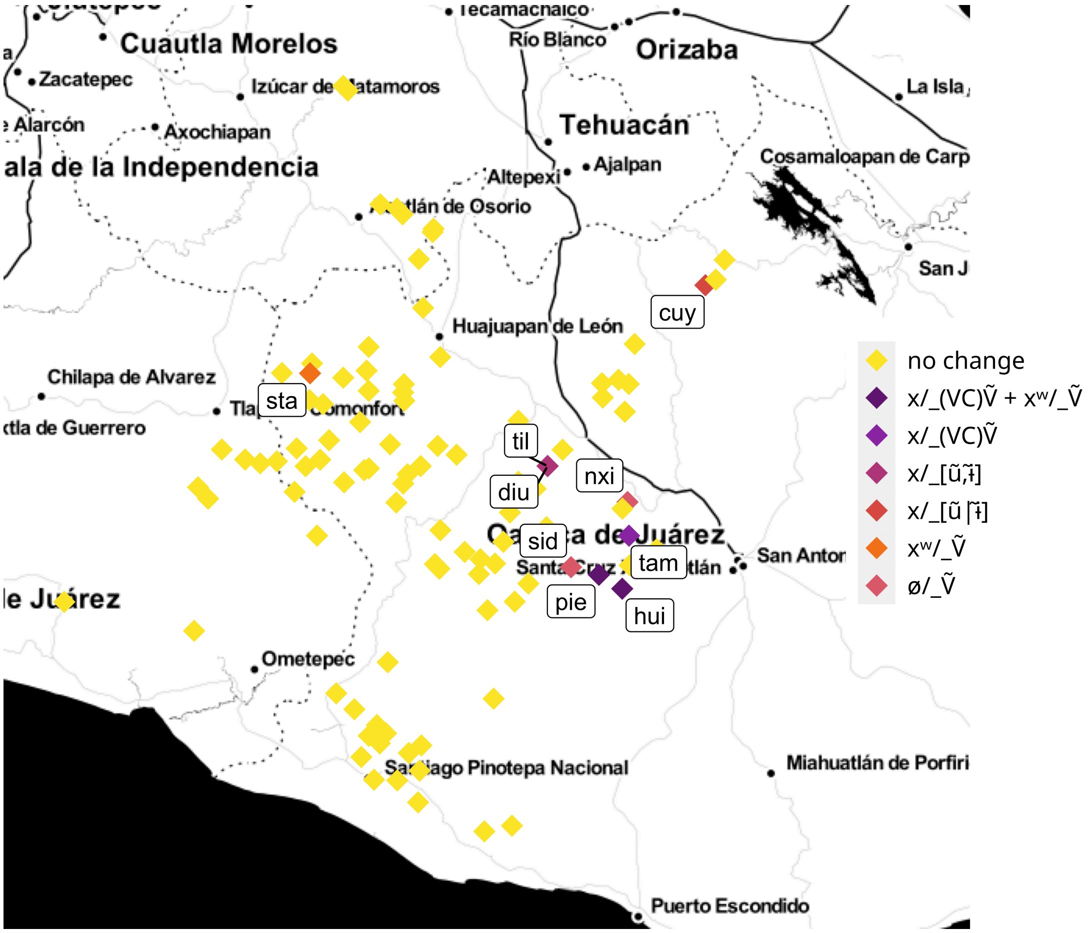


Figure 3.6: Lenition of $*k$ and $*k^w$

from Group 4. The environments for these changes are very specific for some languages, and in some cases, the languages that have undegone the same change are not geographically close; such as the loss of $*k$ in Sindihui and Nuxiño. As far as we can see from our data, these changes do not interact with other changes in these varieties and thus it is not possible to determine any relative chronology that involves them. However, given the limited and somewhat scattered distribution, we infer that the lenition and loss of Proto-Mixtec $*k$ and $*k^w$ are quite recent changes.

3.4.3 Developments of Proto-Mixtec *^{n}d

Proto-Mixtec *^{n}d is overall relatively stable and undergoes few or no changes in most varieties. The changes that do occur can broadly be classified as palatalizations before front and originally central vowels. As opposed to Proto-Mixtec *t, palatalization is most common before Proto-Mixtec *e, followed by original *ɨ, rather than *i. The most common reflexes are $^{n}dʒ$ and $^{n}d^{j}$. Treating ^{n}d as a sequence of *n* + *t*, as Kaufman (in press) does, would not reduce the number of changes we need to posit, since ^{n}d does not participate in the changes identified for *t*. This could be taken as evidence that it should not be treated as a sequence, at least diachronically.

The palatalization changes of Proto-Mixtec *^{n}d exemplify well Josserand's (1983) insight mentioned in Section 4.1: That even though Mixtec languages look quite similar in terms of their phoneme inventory today, they arrived there through a series of sometimes quite different and complex changes. Palatalizations of Proto-Mixtec *^{n}d are not as frequent as those with other dental consonants like Proto-Mixtec *t and *s. In fact, 61 languages in our sample do not exhibit any such changes. Those varieties cover all subgroups and in most of them make up about half of each group. The map in Figure 3.7 reveals complex layers of changes with various combinations of conditioning environments and reflexes resulting from them. The situation is even more complex than the maps suggest, since some of these changes happened before changes to the vowels of the conditioning environment, while others had to have happened after certain vowel changes. We briefly discuss each subgroup in turn.

There are no changes to discuss in Group 1. In Group 2, about half of the languages show palatalization. Palatalization in this group is conditioned by either *i* or *e*, or both, but never by *ɨ*. The varieties exhibiting these changes are all located in the eastern part of the coastal region, and in fact, the split of varieties with and without palatalization of Proto-Mixtec *^{n}d falls exactly long the east/west division of Josserand's (1983) Coast dialect area, except that

Acatepec Mixtec is included in the eastern group, as it is in Auderset et al. (submitted). The most common reflex of palatalization in this group is the palatalized $^{n}d^{j}$ with both conditioning vowels. This reflex is found in all varieties except Acatepec Mixtec, which has ^{n}_dʒ_, and Colorado Mixtec, which lost the plosive element and has a reflex _ɲ_. Both of these could be later developments of $^{n}d^{j}$, which makes it likely that this change took place shortly after the split of the eastern coastal varieties from the western ones.

Most varieties in Group 3 do not exhibit palatalization of Proto-Mixtec *^{n}d. In Yolotepec, we find a change to $^{n}d^{j}$ before *_e_, while in Tlacotepec Mixtec we find a change to ^{n}dz before *_i_. The change in Yolotepec is easier to explain, since some varieties of Group 6 (Chalcatongo, El Grande, Yosondua) spoken not too far away also show palatalization before _e_. The change in Tlacotepec is found in an area where most varieties do not show palatalization. About half the varieties of Group 4 show palatalization of *^{n}d to ^{n}_dʒ_ only before Proto-Mixtec *_i_. Those that do not show this change are found at the edge of the Mixtec region. This could point to this change arising after the split of this group and spreading through contact.

Group 51 within Linkage 5 also shows quite uniform changes. In all varieties of this subgroup, we find palatalization before *_i_ and *_e_ and in all but Apoala Mixtec, the reflexes are ^{n}_dʒ_ before _i_ and ^{n}z before _e_. In Apoala, one could say we find the opposite reflex with respect to the sibilant component, namely ^{n}dz before _i_ and ^{n}_ʒ_ before _e_. Of the other varieties in Linkage 5, only Tidaa Mixtec shows palatalization to ^{n}_dʒ_ only before _i_. Only three varieties in Group 6 show palatalization and only in a very limited environment. In Chalcatongo, El Grande, and Yosondua Mixtec, we find a change to ^{n}_dʒ_ before _e_ in final syllables. This change must have occurred before *_e_ changes to _a_, since we do not find it with words that go back to Proto-Mixtec *_a_.

In Group 7, we find the most variation and it is only here that some languages show palatalization before *_i_, *_ɨ_ and *_e_. We discuss each lower level group separately, since it is difficult to make generalizations across this large group. There are two coherent groupings

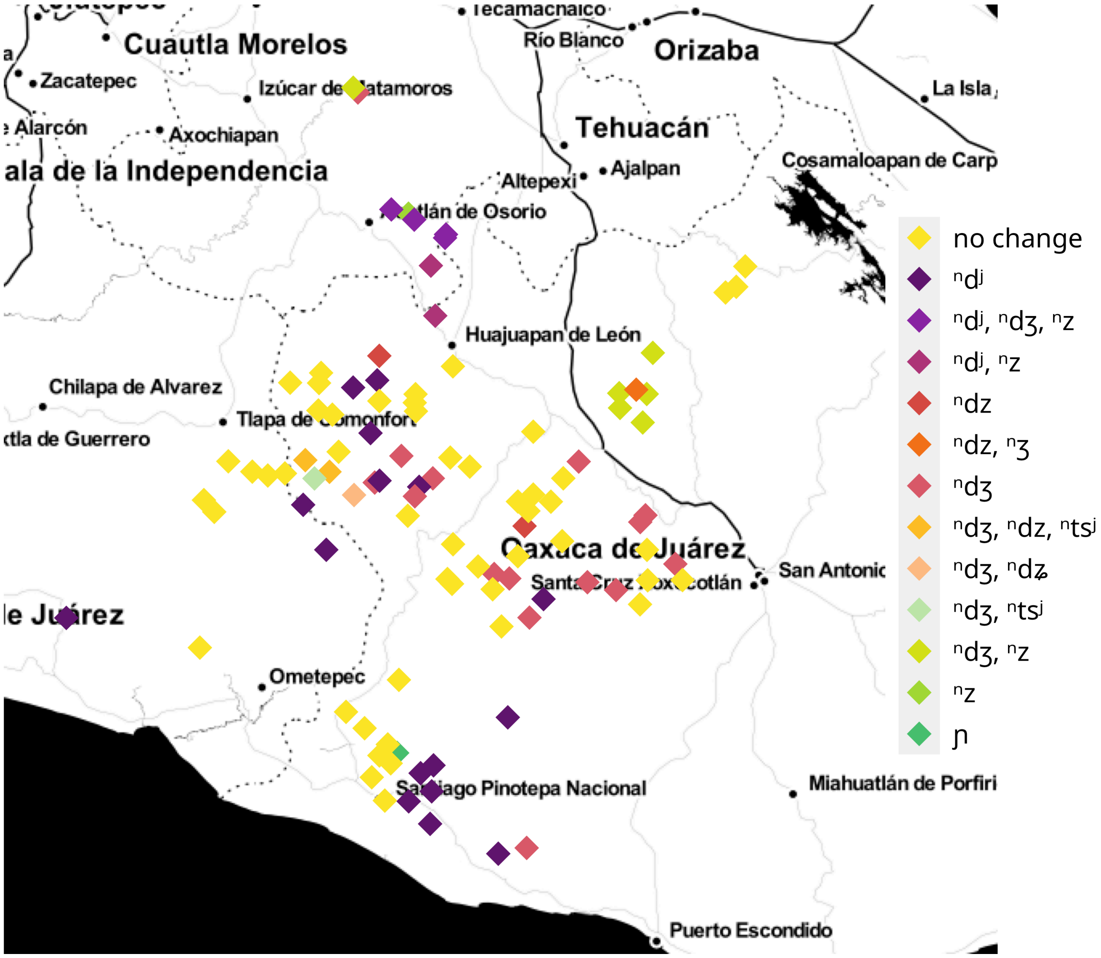


Figure 3.7: Reflexes of palatalalizations in Proto-Mixtec *^{n}d

that exhibit no palatalization: Linkage 75 located at the western edge of this area and subgroup 731 located mostly in Guerrero. Tepango in Group 731 is an exception, changing to $^{n}d^{j}$ before e in final syllables. The rest of Group 73 varieties, except for Progreso Mixtec, exhibit palatalization and fall into two sets based on the conditioning environment: In the border region with Guerrero, three geographically close varieties (Coicoyan, Metlatonoc, and El Jicaral) have palatalization to $^{n}d^{j}$ only before *e, while a bit further east the three closely related varieties of Peras, Piedra Azul, and San Marcos la Flor show palatalization before all three vowels with a different reflex before each vowel. These latter varieties are very close geographically to those of the small Group 72, which shows palatalization before *i, *ɨ and *e as well, although the exact reflexes vary. In Group 71, all varieties except Juxtlahuaca and Yucunicoco, geographically located at opposite ends of this group, have palatalization before *i and *e. All but the variety of San Juan Mixtepec, which has $^{n}d^{j}$, show a uniform reflex ^{n}dʒ. In Juxtlahuaca Mixtec, palatalization to $^{n}d^{j}$ takes place before *i and *ɨ, but not e. Yucunicoco exhibits no such change at all. Group 74 also shows mixed reflexes, but in all languages of the group palatalization happens before *ɨ. Group 76 is again split, with two varieties (San Jorge Nuchita and San Sebastian del Monte) showing palatalization before *e and one before *e and *ɨ. The rest of the varieties show no change.

The only change affecting Proto-Mixtec *^{n}d not yet discussed is confined to two varieties located far from each other and can thus probably be seen as late and independent innovations: Zapotitlan (Group 74) and Nuxaa (Group 4) Mixtec show palatalization to $^{n}d^{j}$ before *u.

3.4.4 Developments of Proto-Mixtec *n

Proto-Mixtec *n undergoes almost no changes and the two that we identify are infrequent. A set of coastal Group 2 varieties and Tepango, which is also spoken near the coast but be-

Table 3.10: Proto-forms and reflexes that illustrate the chronology of n-palatalization

	EIGHT	DOG	Chronology	
Proto-Mixtec	*_o n e_	*_i n a_		
SantiagoJamiltepecMixtec	_u_ ^{1} _ɲ a_ ^{1}	_i_ ^{1} _n a_ ^{1}	1	*n >ɲ/_e
SanAgustinChayucoMixtec	_u ɲ a_	_i n a_	2	*e >a

longs to Group 73, show palatalization of *_n_ before *_e_, cf. Figure 3.8. This change must have happened before *_e_ changed to _a_, since palatalization is only found with _a_ from *_e_ and not from *_a_, as shown in Table 3.10

A few varieties of Group 7, but of various subgroups within that, exhibit palatalization of *_n_ after *_j_ in the previous syllable in cognate set 469 NET, reconstructed as *_jonoʔ_. Due to lack of another _j_-initial cognate set with _n_, we cannot tell whether the change is triggered solely by the palatal glide (which seems likely) or whether it is also restricted to certain vowel qualities.

3.4.5 Developments of Proto-Mixtec *s

Proto-Mixtec *_s_ has undergone relatively few changes, but some of them are pervasive. The two common ones are an unconditioned change to _ð_ and palatalization to _ʃ_ before front vowels. In a few varieties, Proto-Mixtec *_s_ unconditionally changes to _h_, in others this development is conditioned by specific following vowels. The distribution of these main reflexes of Proto-Mixtec *_s_ are summarized in Figure 3.9. This variable roughly splits the varieties into two main groups along a south-west to north-east axis, with the varieties in the north-east primarily exhibiting unconditioned change to the inter-dental fricative _ð_, while those in the south-western area primarily retain _s_ with back vowels and change to _ʃ_ with front vowels. The unconditioned change to _ð_ covers all of Groups 1 and 4, and lower level groups 51 and 74, all located in the far northern and north-eastern Mixtec region. In the north-eastern region near Group 4, the change also appears in Teita Mixtec (Group 3). In Linkage 5 varieties

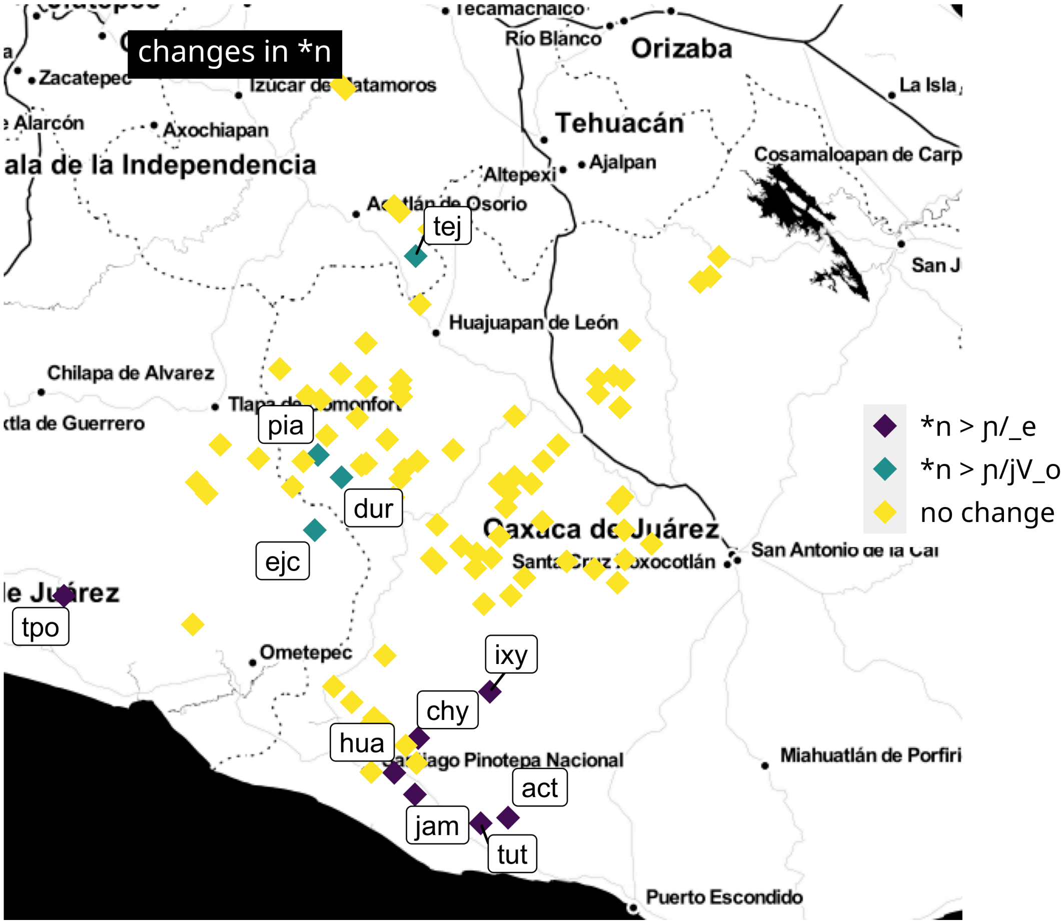


Figure 3.8: Distribution of changes in Proto-Mixtec *n

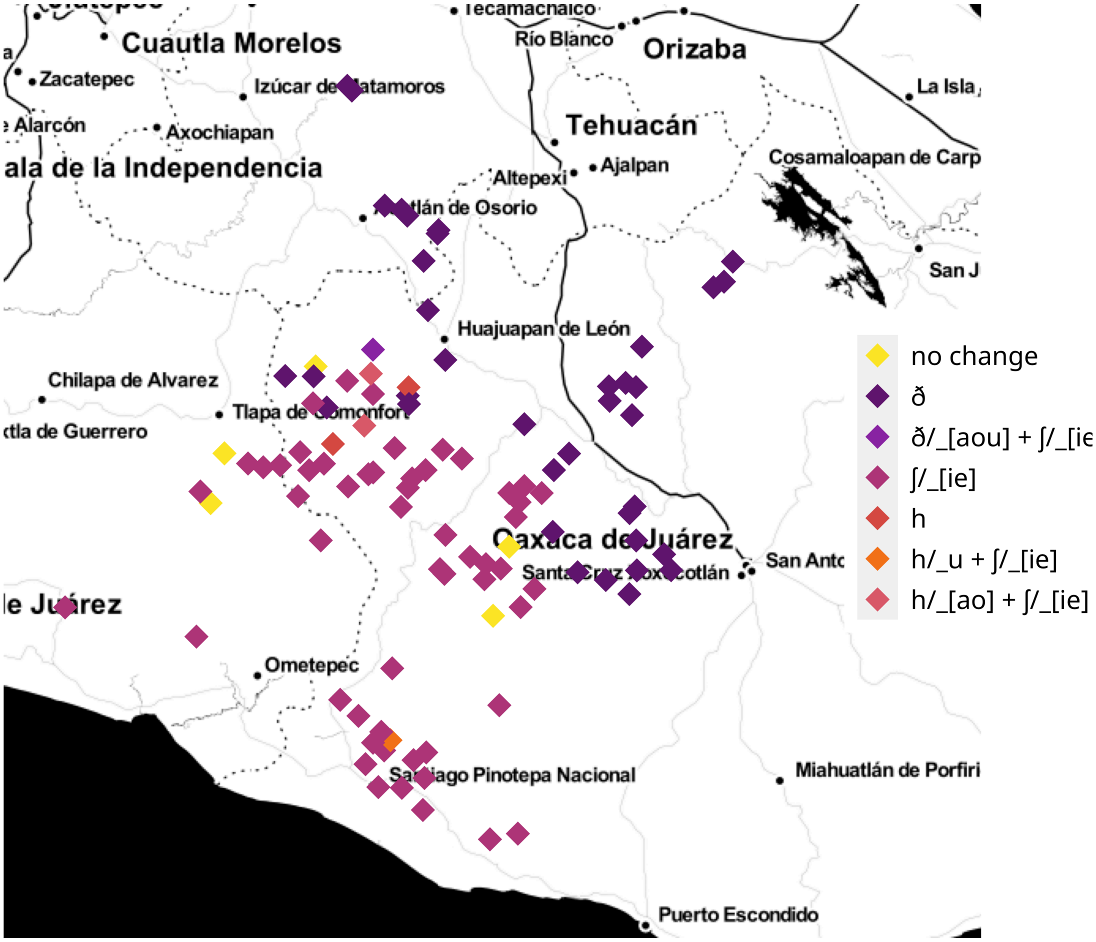


Figure 3.9: Main reflexes of Proto-Mixtec *s

outside of Group 51, we see the same pattern: the three closest to Group 51 – Tidaa, Diuxi, and Tilantongo Mixtec – also show this change.

The other common change, the palatalization of **s* to *ʃ* before front vowels and retention as **s* elsewhere, spans all of Group 2 (with an additional change to *h* before *u* in Colorado Mixtec), all of Group 3 except Teita Mixtec (mentioned above), Ñumi and Yosonama at the southern edge of Linkage 5, all of Group 6 except Molinos and Itundujia Mixtec which retain **s* throughout. Group 7 exhibits a great deal of variation, especially in the Silacayoapam area.

In a few varieties and very specific environments Proto-Mixtec **s* is lost or undergoes metathesis. We can see that the loss always takes place in the context of the front vowel

Table 3.11: Proto-forms and reflexes that illustrate the chronology of s-palatalization

	AUNT	POT	LARD	RAIN	Chronology
Proto-Mixtec	*s i s i	*k ɨ s ɨ	*s e ʔ ẽ	*s a w i ʔ	
PiedraAzulMixtec	ʃ i ^{3} ʃ i ^{3}	k i ^{1} s i ^{3}	ʃ a ^{15} ʔ ã ^{51}	s a^{h} ^{1} β i ^{15}	1 *s >ʃ/_i,e
SantiagoYosonduaMixtec	ʃ i ^{3} ʃ i ^{3}	k ɨ ^{1} s ɨ ^{1}	ʃ a ^{1} ʔ ã ^{1}	s a ^{3} u ^{1}	2 *e >a | *ɨ >i

Table 3.12: Two examples of the different orderings of palatalization and metathesis

	NOSE	HAWK	Chronology
PMx	*s i t ĩ ʔ	*s i ʔ j ã	
SantaLuciaMonteverdeMixtec	i s t^{n} ĩ	ʃ i ʔ ɲ ã	1 methathesis; 2 palatalization
SantaMariaYucunicocoMixtec	i ^{1} ʃ t ĩ ^{5}	ʃ i ^{5} ʔ ɲ ã ^{5}	1 palatalization; 2 methasesis

i, which is also the only vowel triggering metathesis. Interestingly, the metathesis in some languages takes place before palatalization of **s* to *i*, while in others the reverse is true, see Table 3.12. This adds to the impression that these are parallel, possibly sporadic innovations.

3.4.6 Developments of Proto-Mixtec *x

According to Mak & Longacre (1960:33), the sound that Longacre (1957) reconstructs as the Proto-Mixtecan velar fricative **x* unconditionally lenited to **h* in Proto-Mixtec, merging with reflexes of Proto-Mixtecan **k* preceding **i*. This analysis requires positing subsequent changes of fortition to **tʃ* or **ts* in some Mixtec varieties, and back to **x* in others. We follow Josserand (1983) and Kaufman (in press) in reconstructing Proto-Mixtec **x*. This consonant changes to another consonant, unconditionally, in just under half of the varieties. The most frequent reflex is *ʃ*, followed by *s*, but we also find affricate reflexes *ts* and *tʃ*, see Figure 3.10. About another half of the languages exhibit various changes from Proto-Mixtec **x* conditioned by often very specific vowel environments, which we discuss in more detail below. Finally, Proto-Mixtec **x* is fully retained as such in a handful of varieties located in the center of the Mixtec region.

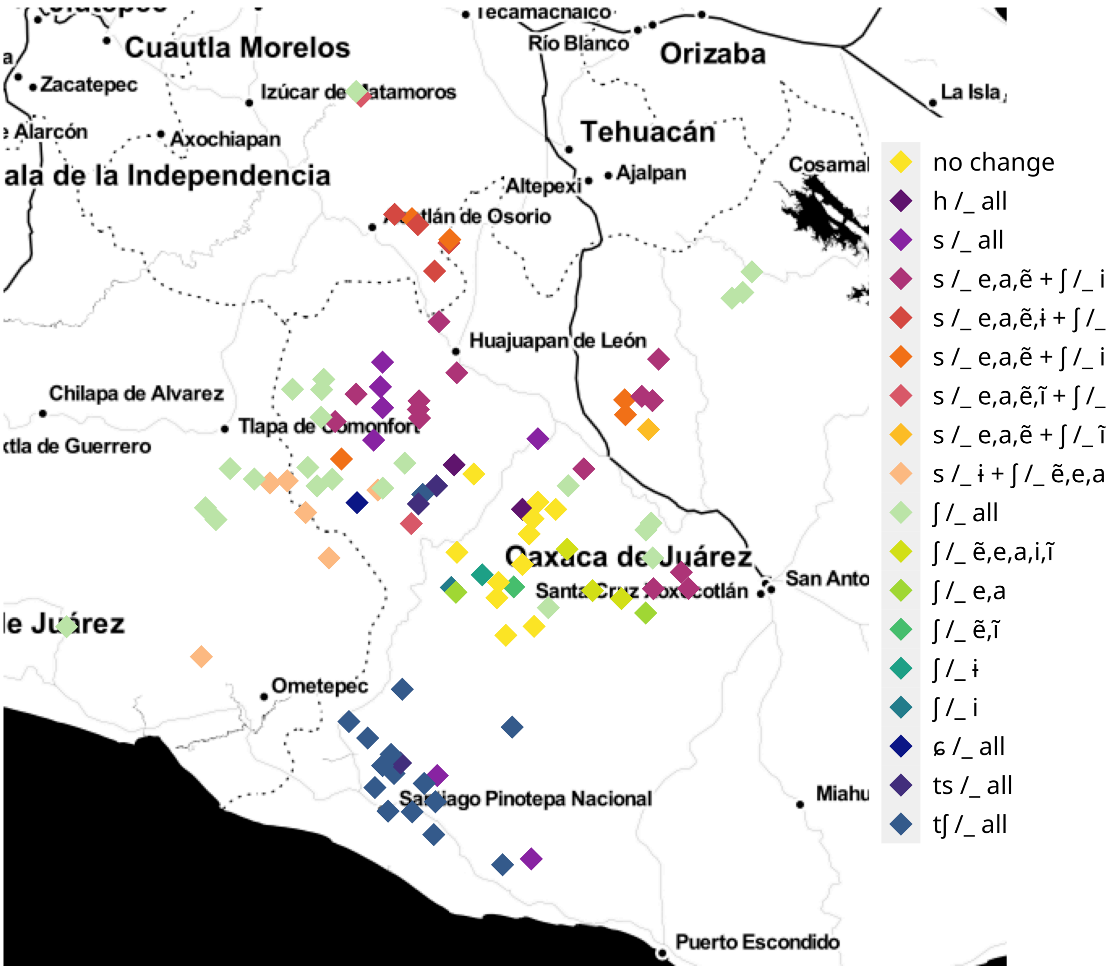


Figure 3.10: Selected changes in Proto-Mixtec *x

In Group 2 almost all languages have a reflex _tʃ_ throughout, while Colorado Mixtec has _ts_. Affricate reflexes – both _ts_ and _tʃ_ are exclusively found in this group and in Group 71 (Mixtepec). The two exceptions to this in Group 2 are Acatepec, which has a reflex _s_ throughout, and Chicahua which has a reflex _ʃ_ before _i_ and _s_ otherwise. In Acatepec, this leads to an almost complete merger of Proto-Mixtec *_s_ and *_x_ (except before _i_, where reflexes of the former but not the latter are palatalized to _ʃ_). A global reflex _h_ is found only Magdalena Peñasco (Group 3) and Yosoñama (Linkage 5). Closely related languages show retention of _x_ for the most part.

An unconditioned change from *_x_ to _ɕ_ is found only in Duraznos Mixtec (Group 72). Group 7 shows no retention of *_x_, but instead various patterns of _s_ and _ʃ_ reflexes – apart from the already discussed Group 71. Group 73 is split between varieties that have a global _ʃ_-reflex and those that have _s_ before _ɨ_ and _ʃ_ otherwise. Group 74 predominantly shows _ʃ_ before _i_ and _s_ otherwise. In Linkage 75, we find another slight variation from that, such that these varieties exhibit _ʃ_ before _i_ and _ɨ_ and _s_ otherwise. This distribution is also found in two varieties of Group 76, while the other varieties in this group either have a global reflex _ʃ_ or _s_. There seems to be a geographic split, with varieties in the north showing more _s_-reflexes and those in the south displaying more _ʃ_-reflexes. A plausible explanation for this larger-scale geographic split with many slightly different conditioning environments would be that proto-Group 7 had _ʃ_ before _i_ and _s_ otherwise. Subsequently, _s_-reflexes started to spread in the north, while _ʃ_ started appearing in more environments in the southern part of this region.

Proto-Mixtec *_x_ is sometimes lost initially or between front vowel _i_ and in one instance also before _ɨ_. Loss of *_x_ is rare and only found in a few varieties of Group 4 and 6, as well as Coatzospam from Group 1 and Teita from Group 3. The latter and the varieties of Groups 4 are geographically quite close to each other, but otherwise it is difficult to explain the distribution of this change.

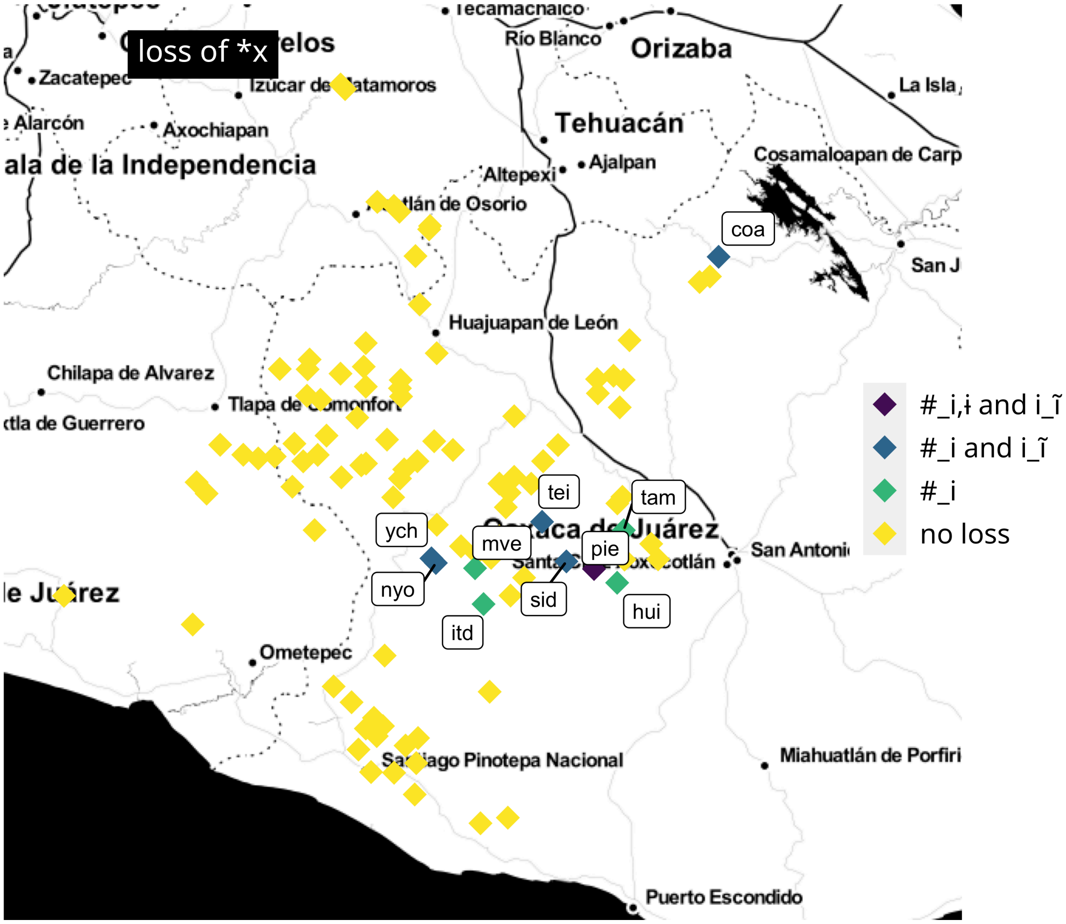


Figure 3.11: Loss of Proto-Mixtec *x

3.4.7 Developments of Proto-Mixtec *j

Proto-Mixtec **j* is the consonant that has undergone perhaps the most changes. There is no variety in which it is retained in all contexts, and it is often found as (part of) the conditioning environment of other changes, requiring relative chronologies in some cases. Proto-Mixtec **j* has only three reflexes: it is either unchanged, lost, or turns into the corresponding palatal nasal *ɲ*. There is one change which we did not code for because it applies to all Mixtec languages without exception: immediately preceding nasal vowels, Proto-Mixtec **j* turns into the corresponding palatal nasal *ɲ*. This change thus must be very old, and given how pervasive it is, it is likely that *ɲ* was an allophone of **j* already in Proto-Mixtec. Subsequent developments then led to its phonologization in most varieties. In some varieties, nasality also affected **j* before *n* in a following syllable, or through an intervening *w*, fricative, or even plosive (see Figure 3.12). These developments are rare, however, and must have taken place after the break up of Proto-Mixtec. Most frequently, we find a change to *ɲ* before a nasal or *w* (with or without intervening glottal stop) in the following syllable. Other combinations of environments are more geographically restricted. Varieties that show the change even with an intervening *x* are confined to Group 3 and Group 6 around Atatlahuca. The change of **j* to *ɲ* in the context of either a following nasal consonant or the semivowel *w* with intervening glottal stop is particularly frequent in Group 7, but it also shows up in one coastal variety. This change is best described as regressive assimilation, since Proto-Mixtec **w* changes to the nasal *m* in this same context. Given that these changes follow neither a clear geographic pattern, nor line up with dialect areas or subgroups, we propose that they reflect a series of parallel developments emerging and spreading locally.

The loss of Proto-Mixtec **j* is conditioned by vowels, often a combination of specific vowels both before and after **j*. The details across varieties vary so much that we identified a total of 14 fine-grained changes. A detailed description of each of these changes' distribution

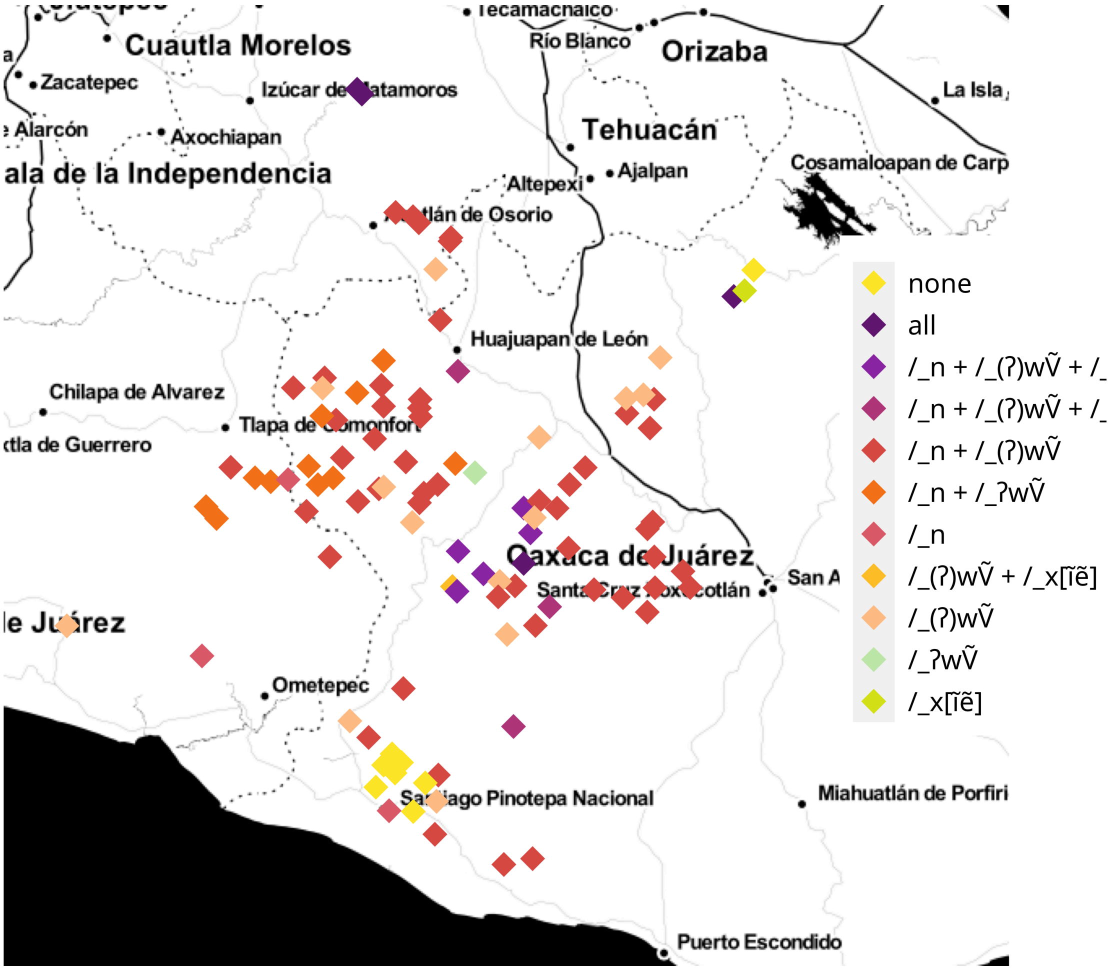


Figure 3.12: Environments conditioning ɲ from *j (besides Ṽ)

lies outside the scope of this paper, but they can be examined in detail in the supplementary materials X. In general, *j is most often lost following front vowels i and e and preceding u. All languages in our sample exhibit at least one of the 14 identified changes. In the Mixtepec area, all varieties of Group 71 except for Yucunicoco have lost *j in 80% of the contexts in the identified changes. Additionally, much of Group 7, which corresponds roughly to the Mixteca Baja geographical area, exhibits relatively high rates of loss of *j. Varieties that are conservative with respect to retaining *j are found predominantly in the south-eastern area. This geographic split is also evident in Group 2 on the coast, where the eastern varieties are more conservative with respect to *j than the western ones.

3.4.8 Developments of Proto-Mixtec *w

The Proto-Mixtec semivowel *w undergoes many changes, and like its palatal counterpart, there are only three reflexes: retention (often as a bilabial approximant *β̞)*, loss, or change to the corresponding nasal m. The change to m before nasal vowels is also pervasive, but in Tidaa Mixtec we do not find m at all. In Tilantongo and Diuxi Mixtec (all of Linkage 5), m is quite marginal because *w is lost in many environments in which it otherwise could have nasalized. Less frequently, *w changes to m before the prenasalized stop nd preceded by a glottal stop. This change is found in all of Group 1, almost all of geographically close Group 51, and in all but one variety of Group 71. It also occurred in a handful of varieties of Group 6 and further away along the coast in the eastern part of Group 2. Nasalization of *w is absent from Groups 3 and 4 in the central-eastern area of the Mixteca and in most subgroups of Group 7 (except for the already discussed Group 71). Given the non-contiguous clusters this change exhibits, as shown in Figure 3.14, it is mostly likely that we are looking at parallel innovation with perhaps three centers: one in the north-east around Group 51, one on the coast, and one around the Mixtepec area (Group 71).

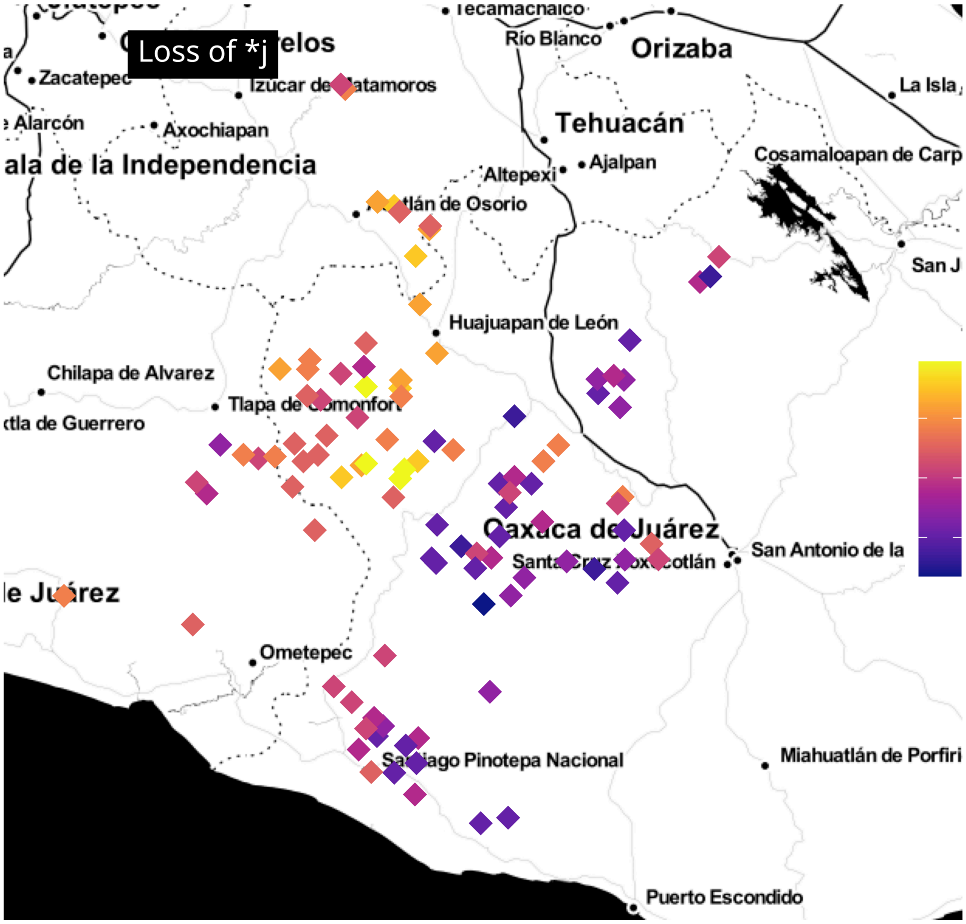


Figure 3.13: Percentage of loss of *j (out of 14 features)

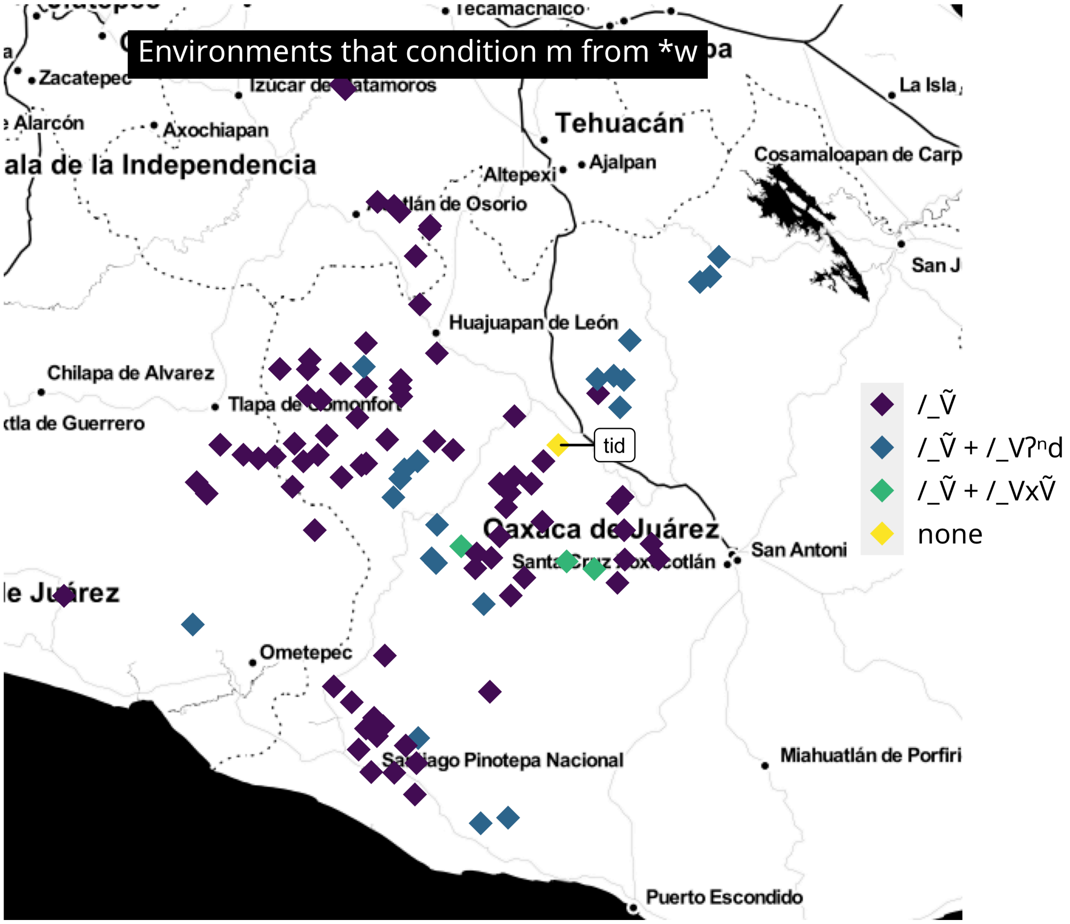


Figure 3.14: Environments that condition m from *w

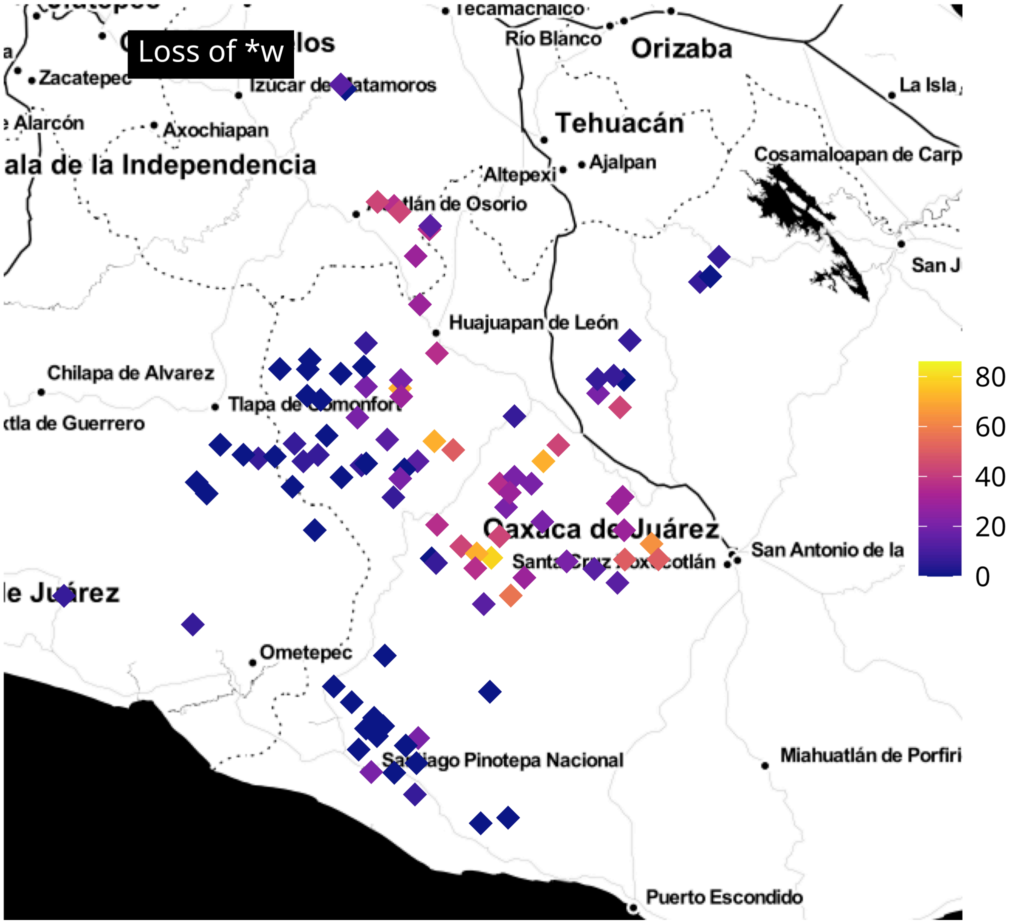


Figure 3.15: Loss of *w

As with **j*, all the other changes identified (16 in total) lead to loss of **w* conditioned by various combinations of preceding and following vowels. A detailed discussion of each of these changes and its distribution lies outside the scope of this study, but we provide a summary of the environments that condition loss as well as an overview in Figure XX. Proto-Mixtec **w* is most commonly lost after the back vowels *u* and *o* and before *i*. Another very wide-spread change is the change to *m* before nasal vowels. Interestingly, many of the varieties that exhibit the most innovation with respect to **j* are quite conservative with respect to **w*, cf. Figure 3.15 and 3.13 in Section 3.4.7.

In about half of the varieties of Group 7, namely those found towards the western part of the Mixtec Baja region, and in most varieties of Group 2, **w* is retained in all environments (except for the change to *m* mentioned above). Another area with little to no loss of **w* is found in the north-east in Group 1 and Group 51.

3.4.9 Developments of the glottal stop

The glottal stop is lost in final position in almost all Mixtec varieties. It is retained only in Zacatepec (Group 2) and Tepango Mixtec in Group 7 (Pankratz & Pike 1967). Conversely, medial glottal stop is preserved in all varieties, with the exception of San Sebastian el Monte Mixtec (Group 7), where it is systematically lost before *w*. This loss of medial glottal stop must be a late innovation. In fact, it could have arisen relatively recently: in this variety, words with intervocalic glottal stop (CVʔV) freely alternate with monosyllabic forms without glottal stop (CVV). It is likely that the systematic loss of the glottal before *w* is related to this variation. In Coatzospam Mixtec, we find an automatic insertion of the glottal stop before word-medial obstruents (Gerfen 1996).

3.5 Vowel changes

In the following sections, we summarize the main developments of each reconstructed Proto-Mixtec vowel and provide maps for the most salient ones. Additional overview maps can be found in the supplementary materials X. The sections are organized roughly from close to open vowels, and within that, following Josserand's (1983) so-called 'inner triangle' vowels (*i, u, a*) and 'outer triangle' vowels (*ɨ, o, e*), because the latter often change into the former.

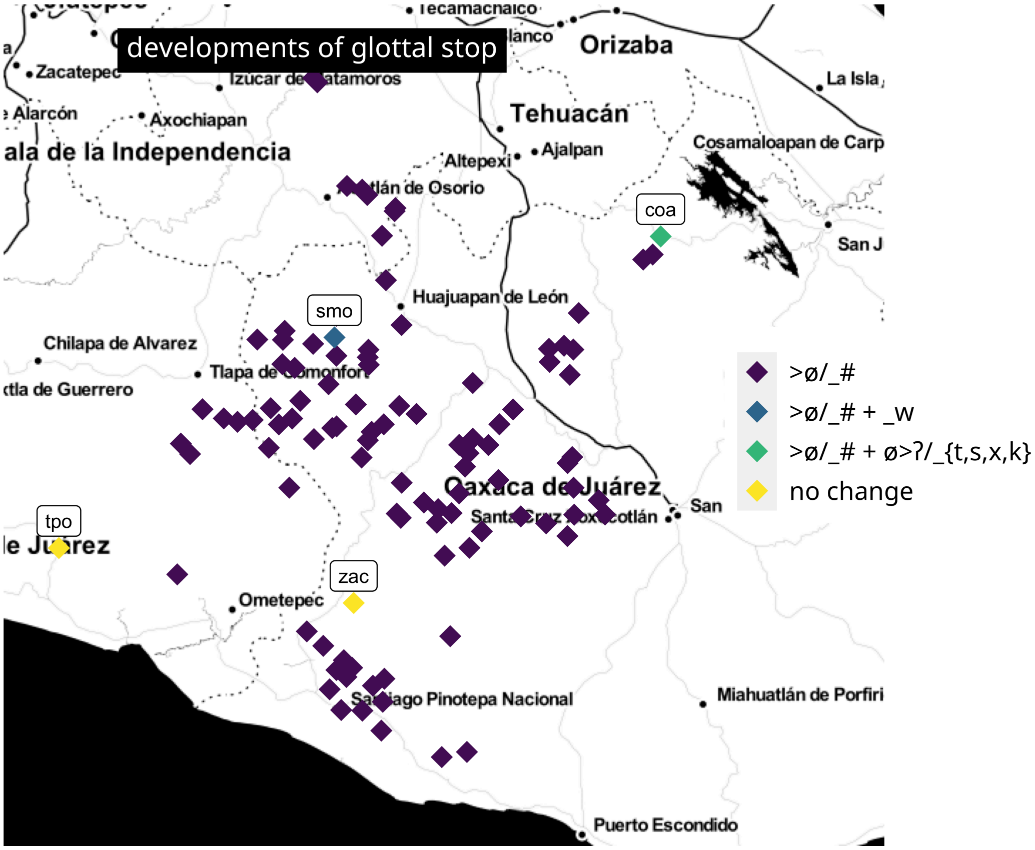


Figure 3.16: Developments of the glottal stop

3.5.1 Developments of Proto-Mixtec *i

Proto-Mixtec **i* has undergone a large number of changes, but in most cases they occurred in only a few varieties. The only frequent change is that to *e* in the context of *e(ʔ)j*. This change is pervasive and appears in all varieties except those of Group 1 and about half of Group 6. The varieties of Group 6 that do not show this change are neither geographically contiguous nor more closely related to each other than to other varieties of Group 6, according to Auderset et al. (submitted).

The other changes mostly involve assimilation of **i* to *u* when *u* occurs in a preceding syllable, and loss of **i* between voiceless consonants. The change of **i* to *u* is – unsurprisingly – most frequent following the labio-velar *k^w or following **w* when there is another **u* in the preceding syllable. These two changes show interesting distributions with respect to larger level groups. Groups 1 and 2 only exhibit the change with the labio-velar, while both changes are present in Groups 3, 4, and 6, and for the majority of Linkage 5. Aside from a few varieties, both changes are absent throughout Group 7. This distribution suggests that this change originated in the Mixteca Alta region, cf. Figure 3.17.

A much rarer change is that of **i* to *u* when a sequence of *a* followed by glottal stop and *w* precedes it. This change is confined to a handful of varieties of Group 6 and a few varieties of Linkage 5, the majority of which also exhibit the previously described changes of **i* to *u*. Changes to other vowels in the same context only appear in one variety each, namely to *o* in Yucuquimi Ocampo Mixtec (Linkage 75) and to *ɨ* in Itundujia Mixtec (Group 6). Changes of **i* to *u* in the context of voiceless plosives are similarly rare and appear only in some varieties of Group 7, namely in about half of Group 7.6 and two varieties of Group 7.3.

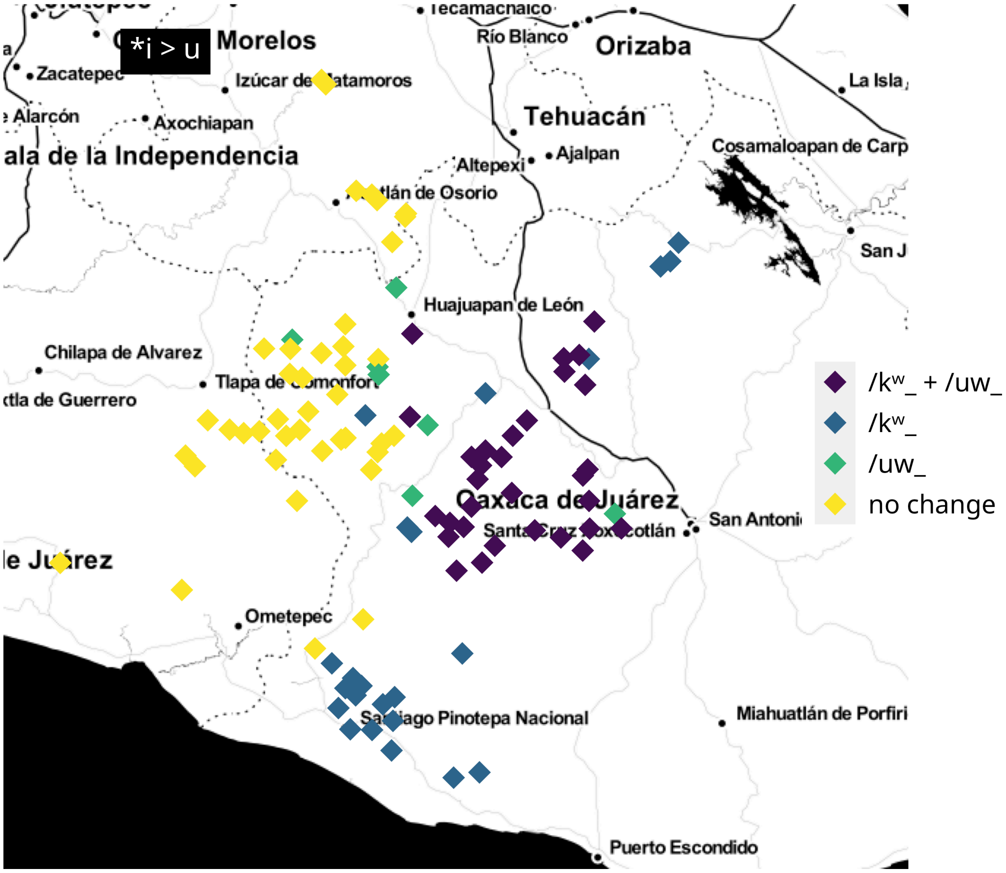


Figure 3.17: Distribution of changes from *i to u in the context of labio-velar plosive and approximant

Table 3.13: Example reflexes that illustrate the ordering of Proto-Mixtec **ɨ* to *i* and palatalizations

	ANIMAL	PATH	Chronology	
Proto-Mixtec	**k ɨ t ɨ ʔ*	**i t i ʔ*	1	*t >tʃ /_i
SanMartinDuraznosMixtec	*k i* ^{3} *tɕ i* ^{1}	*i* ^{3} *tʃ i* ^{1}		*t >ts/_ɨ
PiedraAzulMixtec	*k i* ^{3} *ts i* ^{15}	*i* ^{3} *tʃ i* ^{15}		*t >tɕ/_ɨ
MagdalenaPenascoMixtec	*k i* ^{3} *t i* ^{3}	*i* ^{3} *tʃ i* ^{3}	2	*ɨ >i

3.5.2 Developments of Proto-Mixtec *ɨ

Proto-Mixtec **ɨ* unconditionally merges with *i* in many varieties. This change displays a clear geographic pattern, roughly dividing the Mixtec region along a south-west to north-east axis. The change is found in the north-west, see Figure 3.18. This distribution is congruent with that found by Josserand (1983). This change is completely absent in Group 1 and present in all varieties of Group 7. The rest of the groups show mixed reflexes to different degrees. Four varieties in the border region either exhibit the change only in final syllables or only after **j*. This suggests that the innovation originated in Group 7 and subsequently spread out from there. This would explain the geographic split of Group 2 on the coast, where the change is found only in the varieties spoken in the western part, closer to Group 7. Bradley & Josserand (1982:284) position this change relatively late in their chronology. This fits well with the varied reflexes within most groups, but also with respect to palatalization of consonants. This change must have occurred after the palatalization took place to arrive at the correct reflexes, illustrated in Table 3.13.

Rarely, **ɨ* changes into *u* in the context of **w*. Note that this rare change is correlated with the loss of **w*, that is, all forms that have *u* also exhibit loss of **w*. (The converse, however, is not true: some varieties loose **w* without a vowel change.) This change was identified by Josserand (1983) as sporadic, appearing in only a few geographically and genealogically distant varieties. This is evident in our data as well, although the varieties that exhibit the

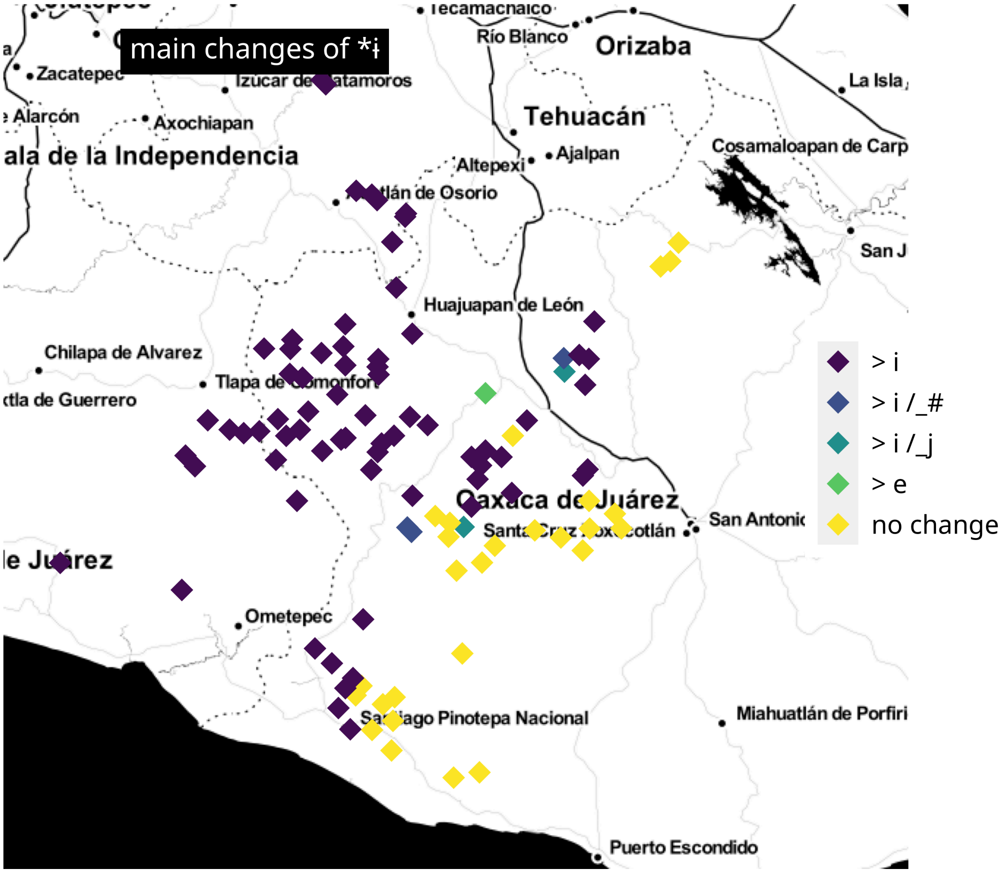


Figure 3.18: Distribution of *ɨ to *i*

change in both the **ɨ* preceding and following **w* are all part of Linkage 5 (Tezoatlan) and the villages are close to each other.

3.5.3 Developments of Proto-Mixtec *u

Of all Proto-Mixtec vowels, we identified the most changes for **u*. However, none of the changes are particularly frequent and often they only occur in a handful of languages and are conditioned by very specific environments. Furthermore, many varieties retain Proto-Mixtec **u* without any changes, cf. Figure 3.19. The changes to Proto-Mixtec **u* seem to have originated around San Andres Yutatio and did not spread widely. This could explain why they are completely absent in Groups 1 and 2, mostly absent in Groups 4 and 6 and Linkage 5, as well as Groups 71, 74 and 76. This together with the limited geographic and genealogical distribution and the large number of changes with specific, non-generalizable environments suggest that Proto-Mixtec **u* is generally stable and the changes have occurred relatively recently.

Most frequently **u* changes to *i* or *o*, often with the same environment resulting in either of these vowels depending on the variety. These changes are most often found in the context of **w* and summarized in Figure 3.20. As can be seen, they are largely confined to varieties of Group 7 in the western part of the Mixteca. However, neither the changes to *i* nor those to *o* line up neatly with subgroups within Group 7, except that both changes are completely absent from Group 71. Changes of **u* to *i* are found throughout Group 76, which displays no changes to *o*. In Linkage 75 all varieties show changes to *i*, but only Yutatio and Yucuquimi also show changes to *o*. In Group 73 about half the varieties show no change at all, while most of the other varieties show both changes. Conversely in Group 74 these changes are absent, except one change in Zapotitlan Palmas in which **w* changes to *o*. The limited distribution and generally poor correlation with subgroups suggest that these changes may be quite re-

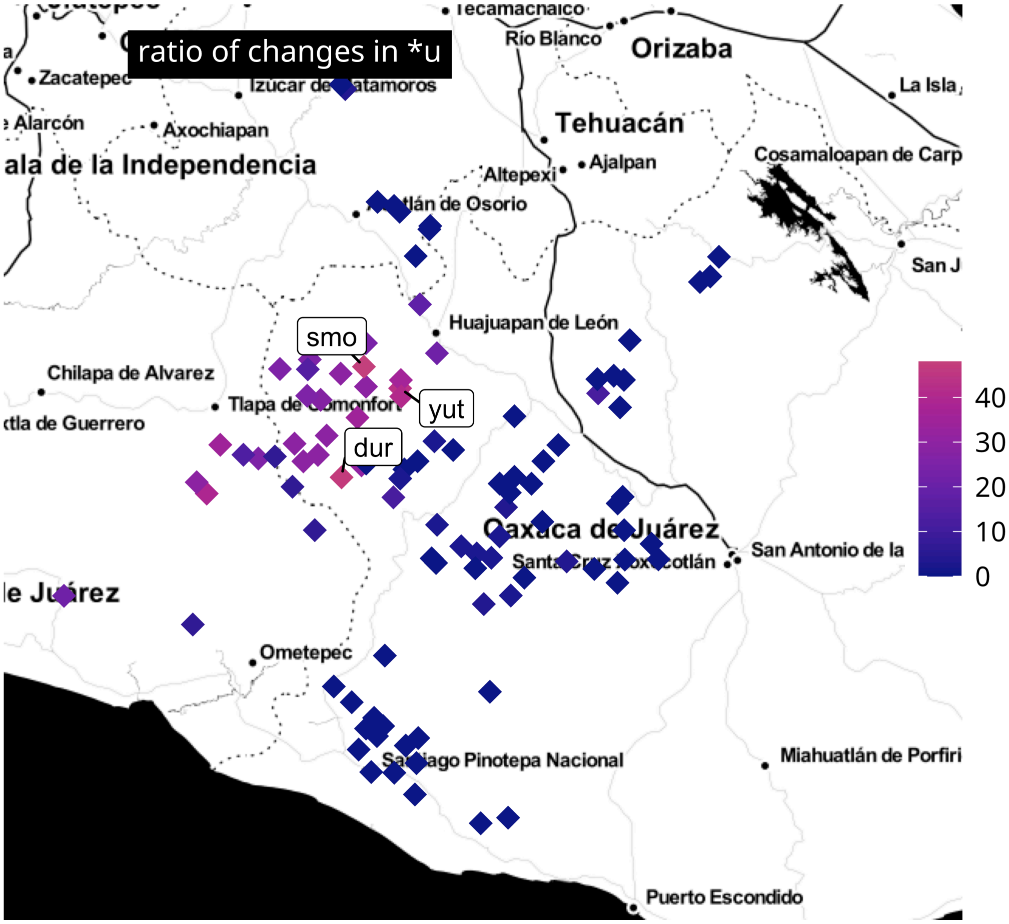


Figure 3.19: Distribution of changes in Proto-Mixtec *u

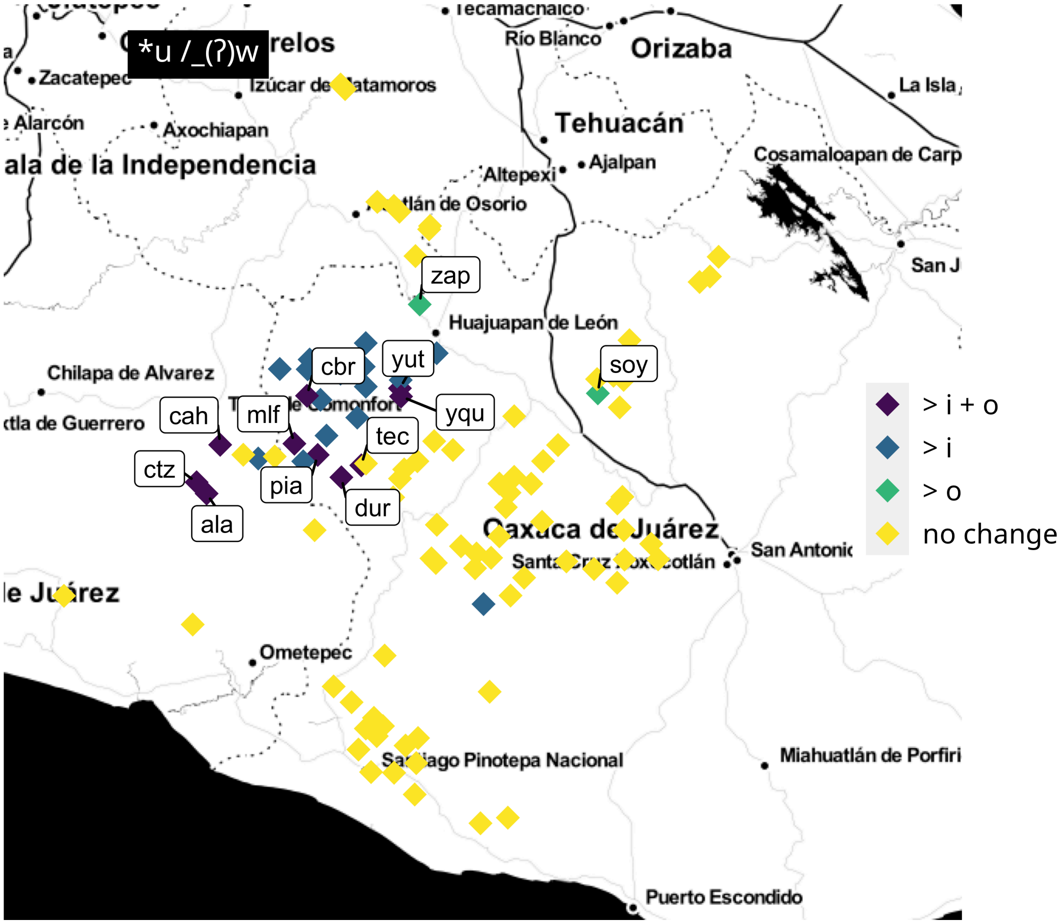


Figure 3.20: Distribution of changes in Proto-Mixtec *u to *i* and/or *o*

cent. Two varieties outside of Group 7 show a change to *i* and *o* respectively: Itundujia and Soyaltepec. However given that they are completely surrounded by varieties that show no change and are located not only far from each other but also far from Group 7 varieties, we view these changes as independent and perhaps sporadic innovations.

An even more limited distribution is found with changes of nasal **ũ* to *õ*. There are two varieties within Group 7 in which **ũ* changes to *õ* everywhere, namely in San Sebastian del Monte (76) and Yutatio (75). The varieties around them show various changes to *õ* in specific environments.

3.5.4 Developments of Proto-Mixtec *o

Proto-Mixtec **o* does not undergo many changes, but the few changes identified are pervasive. In most instances, **o* changes to *u*. The distribution of changes from Proto-Mixtec **o* is quite unlike those of other vowels in that both the most conservative varieties are found in an area in the Mixteca Baja region that is innovative with respect to most other Proto-Mixtec sounds, see Figure 3.21. Conversely, the usually more conservative Alta region shows a high number of changes almost throughout. The varieties of Group 2 are split into an eastern and western group – as seen with other changes – but here the western varieties show a higher number of changes, as seen with other Proto-Mixtec developments.

We have identified changes to Proto-Mixtec **o* before **w*, in the context of the nasal **n*, and with nasalized *õ*. There are no other environments that condition changes to this Proto-Mixtec vowel. Figure 3.22 summarizes the reflexes and conditioning environments with respect to the glide **w*. The change to *u* affects almost all varieties, regardless of the of the vowel in the following syllable. All of Group 1 and three scattered varieties have the change only with *e* or *a* after **w*. Conversely, a set of varieties in Group 7 and one scattered variety in Group 4, Nuxiño, have the change only with *i* after **w*. A handful of varieties in Group 7 have *e* and *i* reflexes, respectively. Almost all of these are found in Group 76, but the change to *i* has also spread to neighboring Ixpantepec Nieves from Group 72.

We discuss the developments of nasal **õ* and **o* in the context of *n* together, since the latter vowels are secondarily nasalized. Figure 3.23 summarizes the environments in which nasal(ized) **o* changes to nasal(ized) *u*. In the majority of languages, the change to *u* happens in all conditioning environments identified. The other combinations of conditioning environments are much rarer and scattered across the Mixteca without clear patterns, except that they are more frequent in Group 7 than any other group and that they divide Group 2 on the coast. In the latter, the languages spoken more to the east exhibit the change to *u* only before

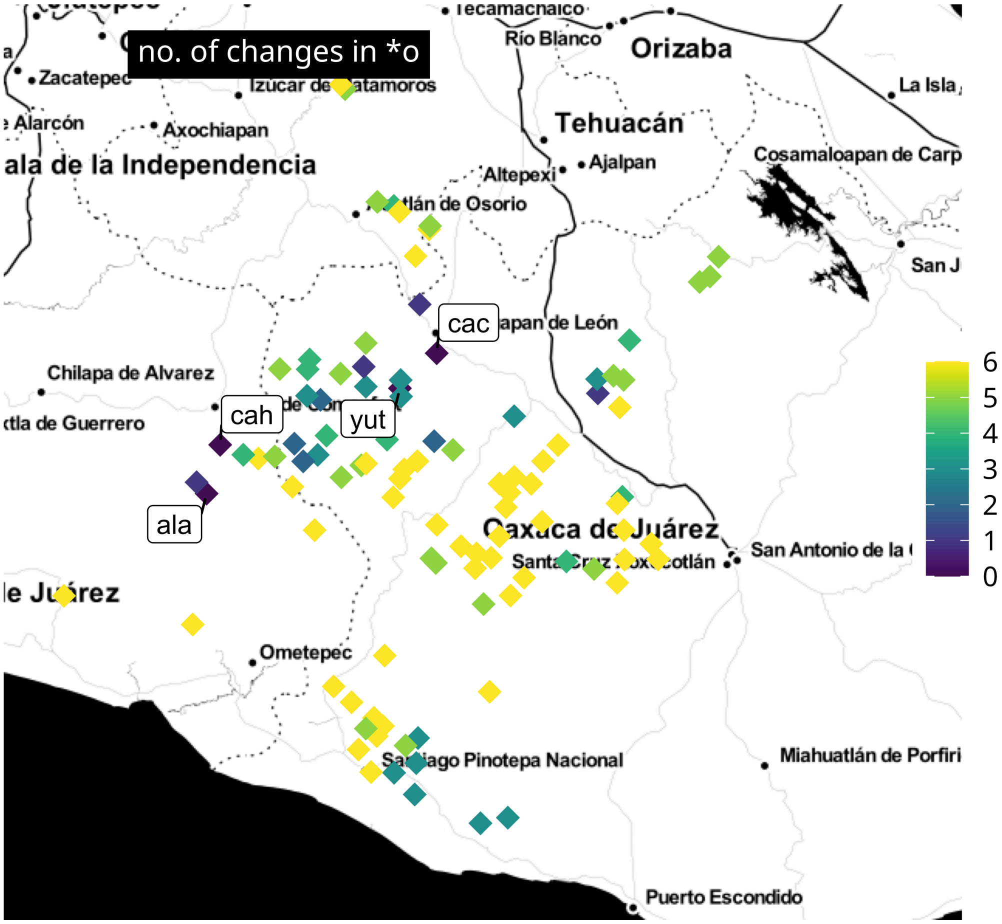


Figure 3.21: Distribution of changes in Proto-Mixtec *o

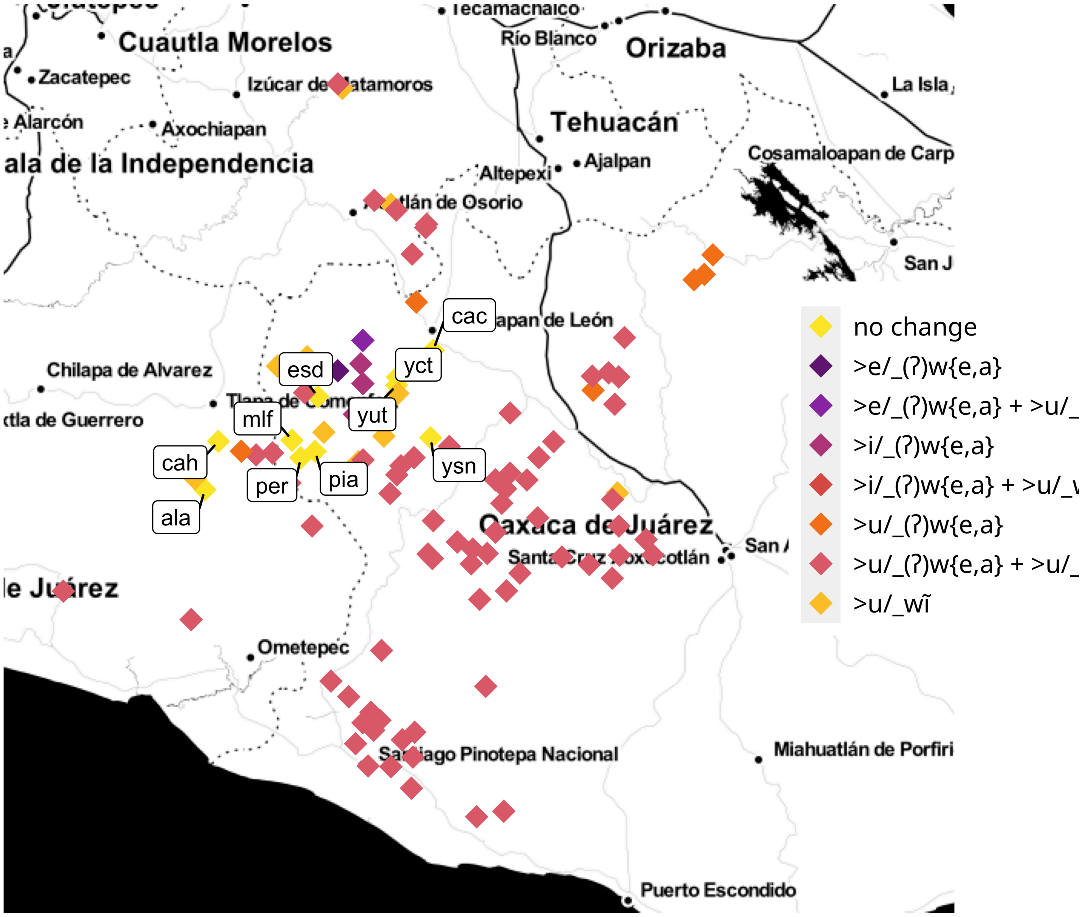


Figure 3.22: Distribution of changes in Proto-Mixtec *o

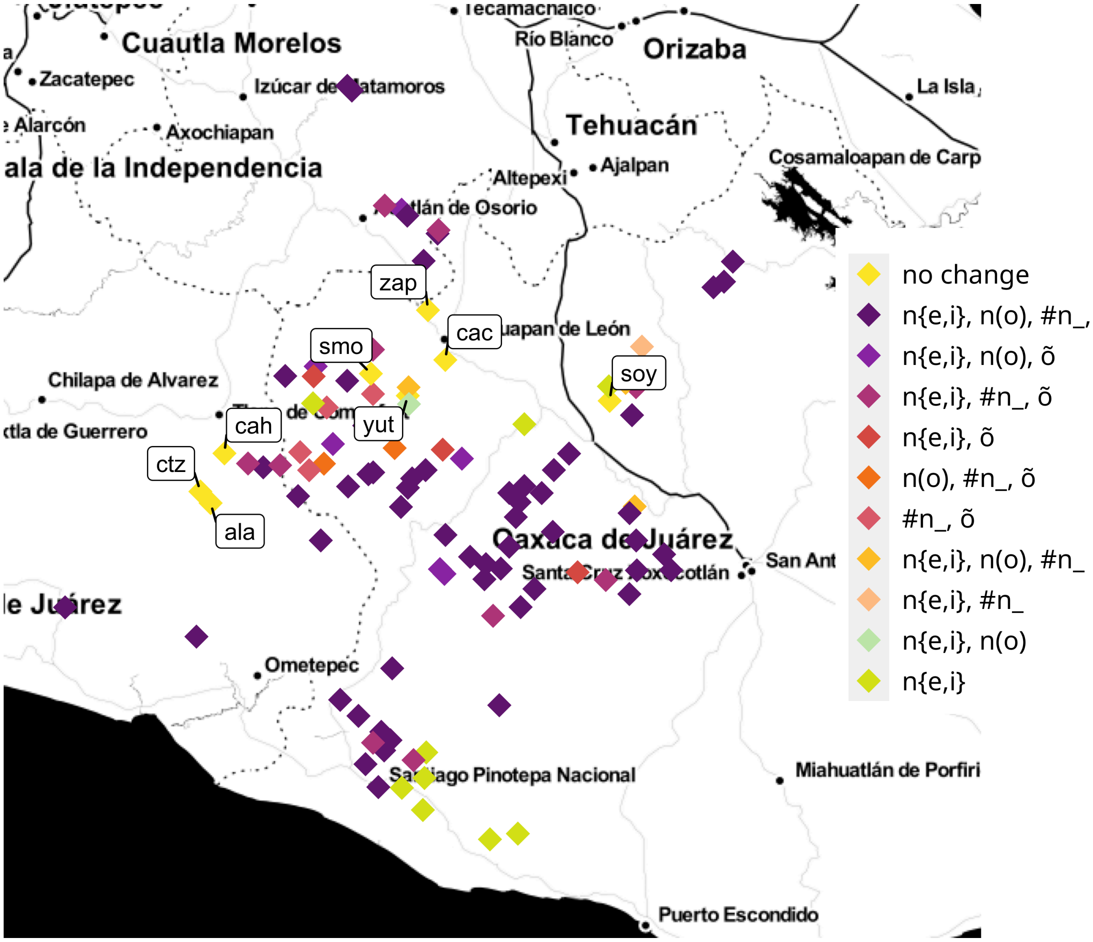


Figure 3.23: Distribution of changes in Proto-Mixtec *o

n with a following front vowel.

3.5.5 Developments of Proto-Mixtec *e

In many varieties, Proto-Mixtec **e* changes to *a* in most contexts and is only preserved in the context of a following glottal stop and the glide *j* (cf. also Josserand 1983:422-448). There are, however, varieties that preserve Proto-Mixtec **e* rather well. Such a widespread retention is found in all of Group 1 and most of Group 4. Conversely, Groups 2 and 7 and Linkage 5 exhibit the most changes (see Figure 3.24). Josserand (1983:424) identified Coatzospam (Group 1) and Itundujia (Group 6) Mixtec as the only varieties that retain **e* in all environments except

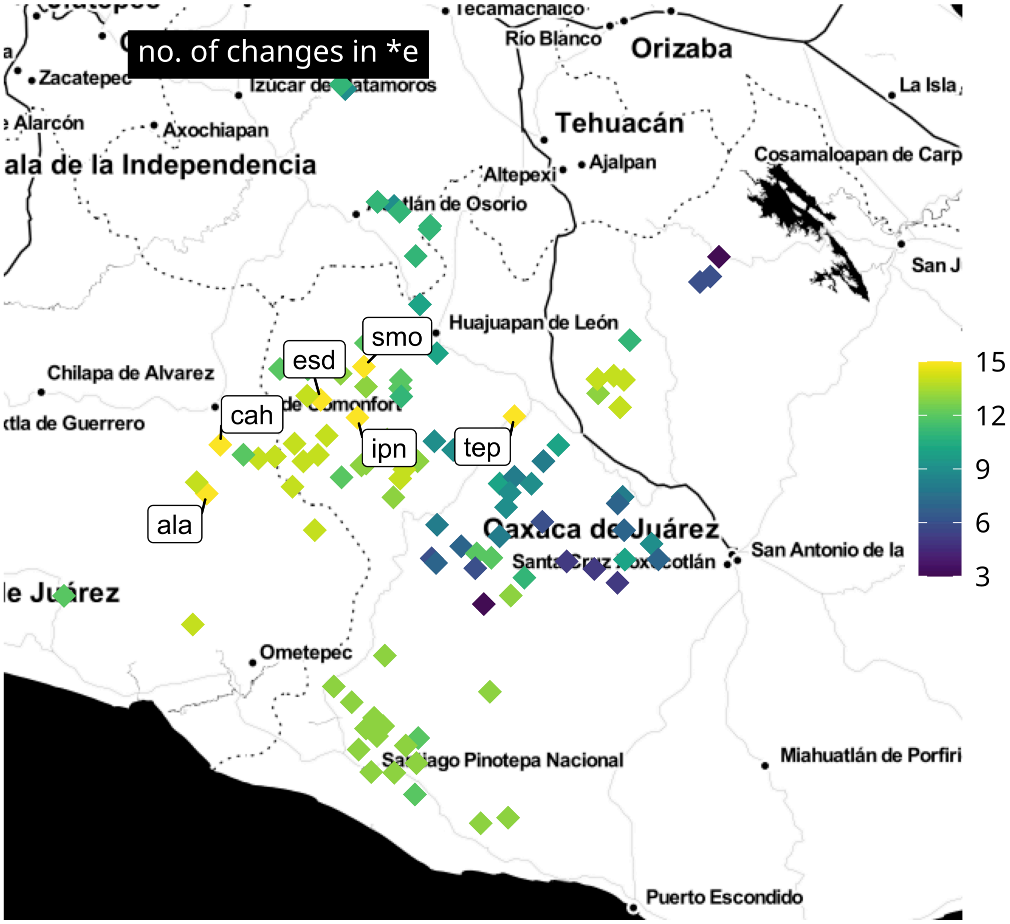


Figure 3.24: Distribution of changes in Proto-Mixtec *e

those with glide *j*. These two varieties are also the ones in our sample that retain **e* in the most contexts.

There is one change that affects all varieties sampled, namely the change of **e* to *a* in the numeral EIGHT (207), reconstructed as **one*. Reconstructing the final vowel as *e* instead of *a* is necessary, because we find palatalization of *n* in exactly those varieties that exhibit it in all cases of *n* followed by *e*, cf. Section 3.4.4. Josserand (1983:423-425) also observed that changes involving Proto-Mixtec **e* are at least partially determined by the class of preceding consonant. She states that **e* remains unchanged most frequently in the context of alveolar consonants. However, our data suggest that nasality and the preceding or following vowel quality are equally important, as the number of varieties that retain **e* does not differ in the context of velars versus alveolars.

3.5.6 Developments of Proto-Mixtec *a

Proto-Mixtec **a* has undergone few changes, as already observed by Josserand (1983:410-421). The only change that is somewhat frequent is the change of **a* to *e* in cognate set 103 (CHILD), where it appears before glottal stop followed by the glide *j* and the vowel *e*. This is probably an assimilatory change triggered either by just the presence of *e* in the following syllable or *e* in combination with *j*. However, we do not have any other proto-form in our sample that has either of these environments, leaving this generalization as an open question for further research. This change affects all of Group 2 and Group 6 and parts of Groups 3 and 4, although with no clear geographic or genealogical pattern in the latter. Within the large Group 7, this change is reflected in all of Group 7.1, 7.4, and 7.5, while Groups 7.2, 7.3, and 7.6 each only have a few varieties exhibiting the change, again with no clear geographic or genealogical pattern in the latter. A change of **a* to *i* in the same environment occurred only in Group 1 and the coastal variety of Tepango. In the latter, however, this is a later, independent development (Josserand 1983:415).

Since the initial **s* shows no palatalization, the change to *e* must have happened after palatalization took place (S05). It remains an open question whether this change took place before or after changes affecting **ʔje*, a problem already identified and discussed by Josserand (1983:415-416). Table 3.14 gives an overview of different reflexes, where we can see that the reflex of **a* neither strictly correlates with the loss or retention of **j*, nor with the reflex of the final vowel.

There are six varieties that show fronting of **a* to *i* after *j*, even with a preceding *a*. As

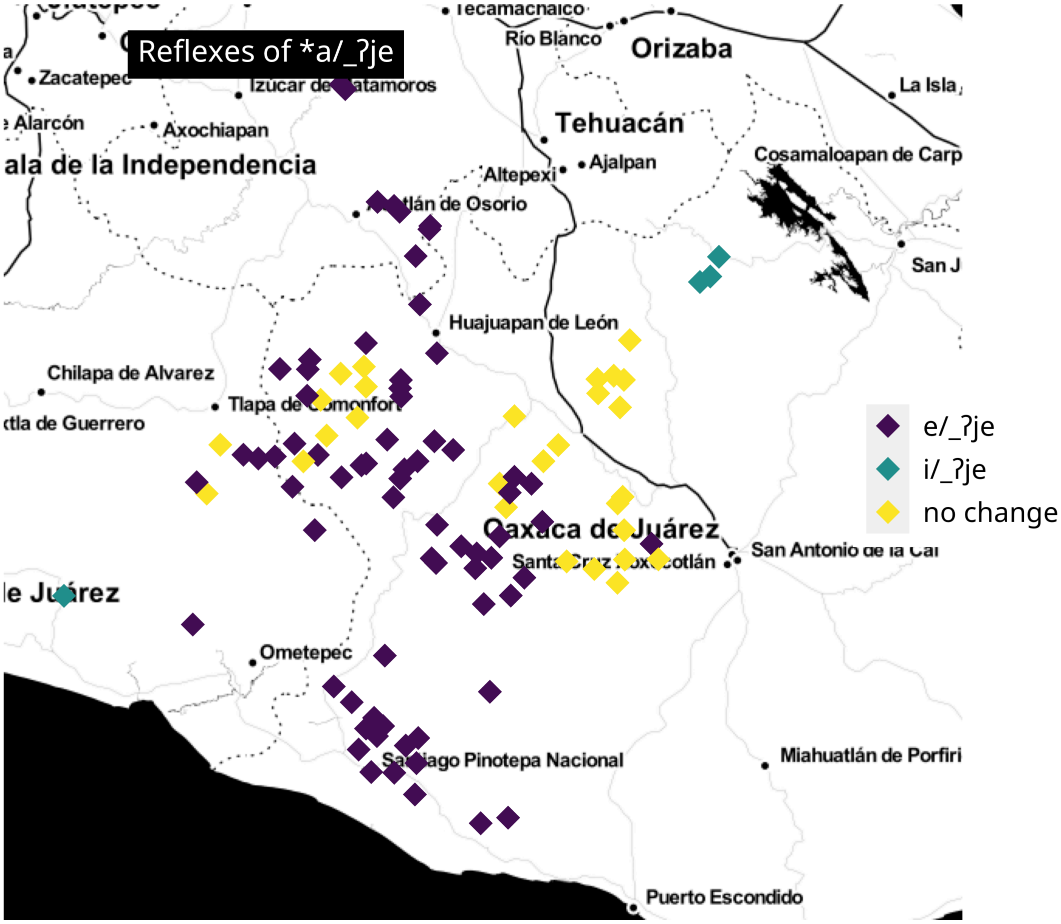


Figure 3.25: Distribution of reflexes from Proto-Mixtec *a before ʔje

Table 3.14: Example reflexes of CHILD (103)

Proto-Mixtec	*s a ʔ j e
AlacatlatzalaMixtec	s a ʔ j a
SanAgustinAtenangoMixtec	h a ʔ j i
SanAndresNuxinoMixtec	ð a ʔ a
SantiagoNuyooMixtec	s e ʔ j a
PinotepaDonLuisMixtec	s e ʔ e
SantiagoTamazolaMixtec	θ e ʔ i
TepangoMixtec	s i ʔ e
SanJuanCoatzospamMixtec	i ʔ ʃ a

with the changes discussed before, we only have one cognate set with this environment in our data, namely LOOSE (407). The change covers four varieties of Group 7.6 (Atenango, Ahuehuetitlan, Nuchita, Villahermosa), one variety of Group 7.3 (Coicoyan) and one variety of Linkage 7.5 (Cacaloxtepec). The latter is also the only variety in which the preceding **a* changes to *e*. These varieties are neither particularly closely related nor geographically contiguous, suggesting late and independent developments.

There is a set of varieties of Group 7.3 that show an epenthetic initial *j* in cognate set 755 (TASTY). The only other *a*-initial proto-form – 840 WHEN **awã* – does not show this epenthetic *j* in any variety. This, together with a similar change before initial **i* followed by a dental plosive (see Section 3.5.1) suggests that conditioning factor here is the **s*. Note that the latter change only affects a subset of the varieties that show epenthetic *j* before **a*, indicating perhaps that this change took place earlier.

The two changes to mid vowels in the context of **w* are only found in Yutatio Mixtec (75), which generally exhibits many vowel changes, and one in Coicoyan Mixtec (73), which could be sporadic. The change to *o* in the context of glottal stop and **w* (A06) is perfectly correlated with the loss of **w* in the same context (W09) indicating that the vowel change probably happened at the same time due to the loss of the semivowel.

3.5.7 Changes in nasalization

The loss of nasalization is relatively common in Mixtec languages, but it usually only affects selected lexemes and not always the same ones across varieties. This is thus not a regular sound change in some varieties that we can code for as we did with other changes. A detailed study of the loss of nasality in each variety would be worthwhile for future research, to investigate whether or not the loss permeates the lexicon in a similar way in different varieties. Here, we summarize our main observation from the data as to whether vowel nasality

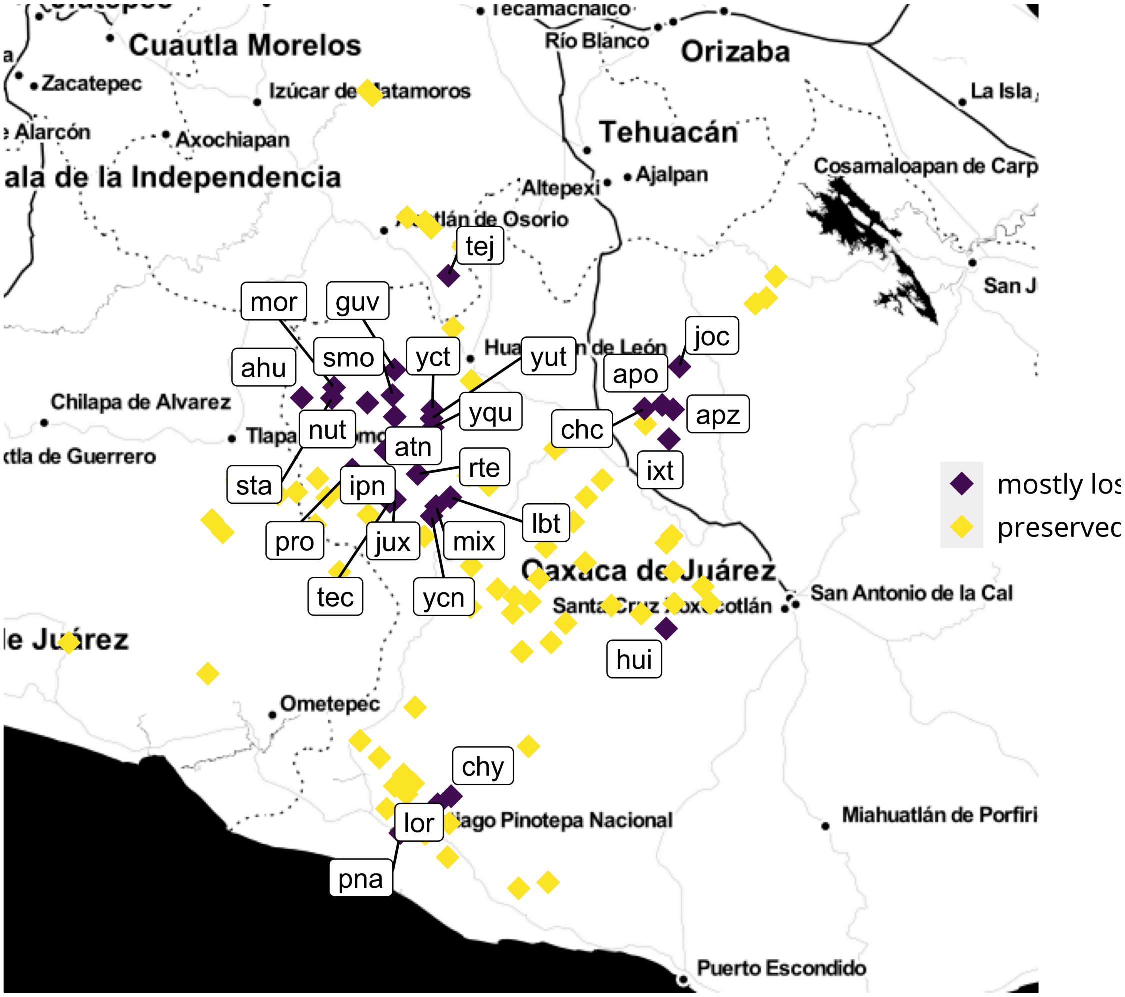


Figure 3.26: General developments of nasalization

is generally preserved or generally lost in a variety. The results of this summary are presented in Figure 3.26. Vowel nasality is lost to a great extent in about a third of the languages. One cluster of such languages is found in the north-west where it covers all but one variety of Groups 71 and 76 and Linkage 75, as well as a few geographically adjacent varieties of Groups 72 (Ixpantepec Nieves, Tecomaxtlahuaca) and 73 (Progreso). Another smaller cluster is found in the north-east where it covers all but one variety of Group 51. We also find it in three centrally located varieties on the coast from Group 2 and in Huitepec from Group 4.

3.6 Correlations between sound changes and implications for subgrouping

In their sound change study, Bradley & Josserand (1982) proposed a relative chronology of changes and based on that and the isogloss maps of the 16 sound changes they discussed, they sketched a diversification scenario for the Mixtec languages. The main point of interest from their proposal relates to the origin of the diversification of the family as such and to the origin of the migration to the coast. Bradley & Josserand (1982) describe the first split to happen after the Proto-Mixtec period as between Soyaltepec (Linkage5, Group 51) and Tilantongo (Linkage 5) versus Teposcolula, which represents the varieties that do not show the innovation. Although this is not stated explicitly, this scenario suggests that the diversification of Mixtec started in the area of the Nochixtlan valley. The first migration to the coast in this scenario takes place in stage 5 (out of 8) from Mixtepec and includes the varieties of Zacatepec, Jicaltepec, and Pinotepa de Don Luis (all Group 2). The second migration to the coast happens in stage 6, again from Mixtepec, but this time only including the variety of Tututepec.

While a complete chronology of all the changes identified lies outside the scope of this study, we can use the overall number of changes as a proxy of how conservative or innovative varieties are. This information is summarized in Figure 3.27 for the total number of changes and in Figure 3.28 and 3.29 for consonants and vowels separately. The most conservative varieties overall are Yucuhiti (Group 6), Itundujia (Group 6), and Coatzospam (Group 1), but there is also a cluster of very conservative varieties on the coast. Conversely, we find the most changes in Ixpantepec Nieves, Santos Reyes Tepejillo, and San Andres Yutatio (all Group 7) and the varieties around them.

Looking at just the changes pertaining to consonants, all of the most conservative varieties are located in the eastern part of the coast around Pinotepa de Don Luis. The most innovative

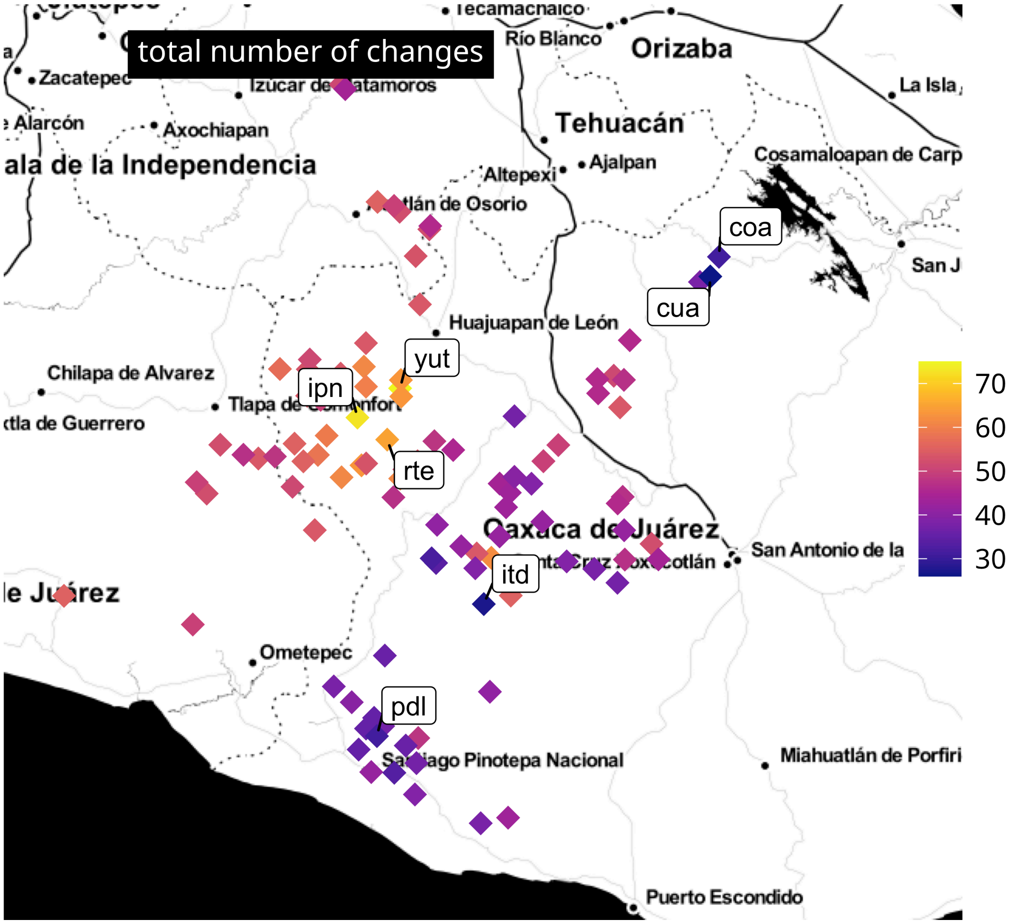


Figure 3.27: Total number of changes per variety with labels for the three varieties with the most and least changes

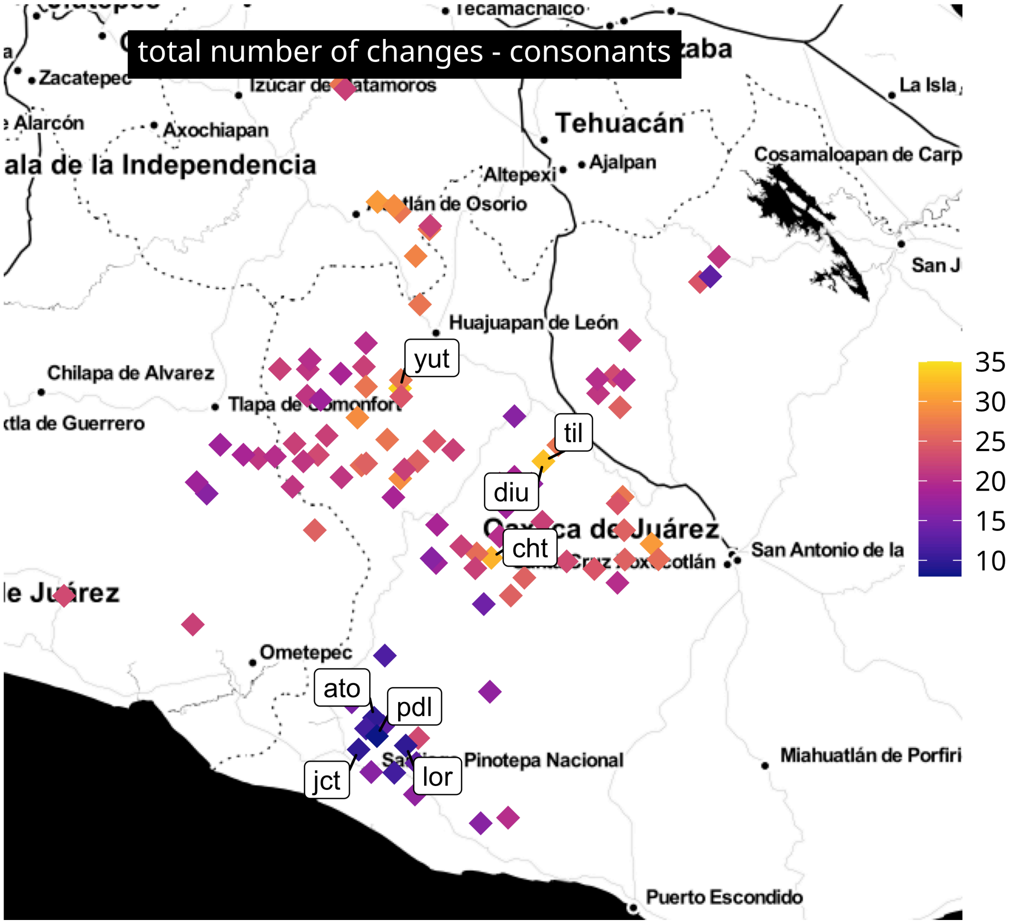


Figure 3.28: Total number of consonant changes per variety with labels for the three varieties with the most and least changes

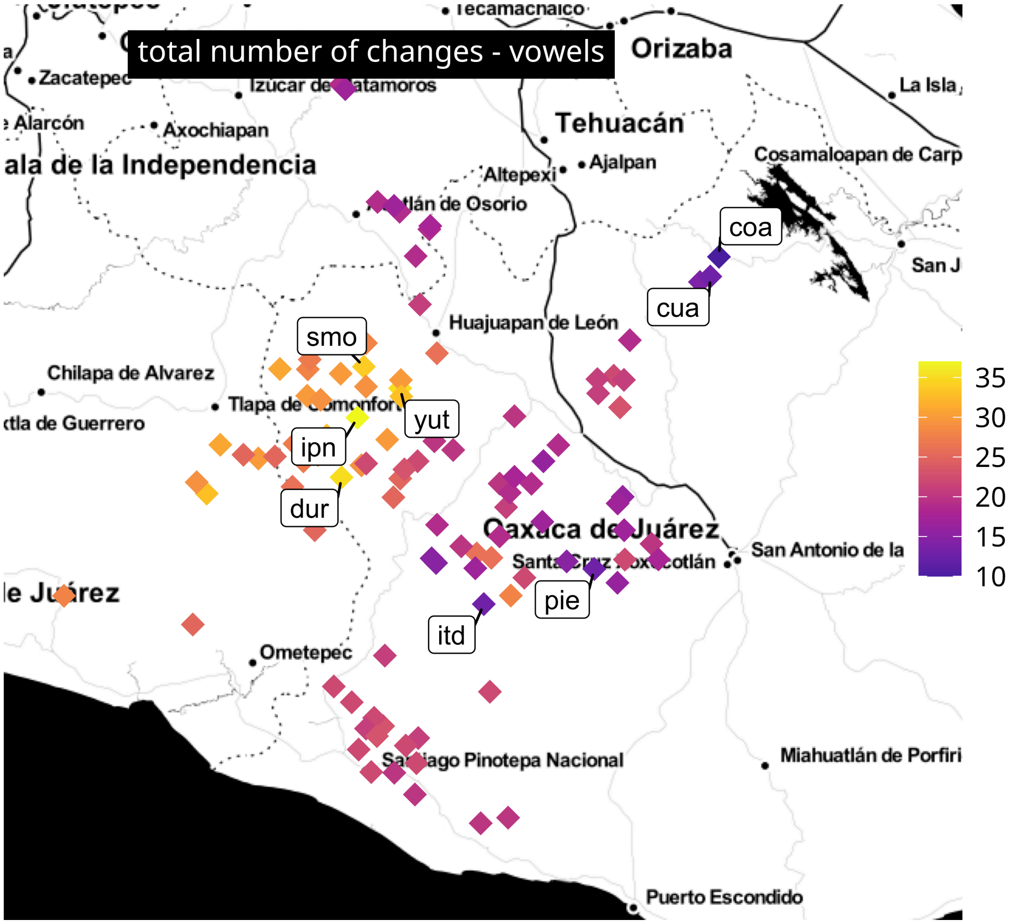


Figure 3.29: Total number of vowel changes per variety with labels for the three varieties with the most and least changes

varieties are here not only found in Group 7, but also in Linkage 5 with Diuxi and Tilantongo. Generally, looking at the just the consonantal changes, only the coast is clearly set apart from the rest of the Mixteca by virtue of being so conservative. There also seems to be a tendency for varieties in the south of both the Baja and the Alta to exhibit relatively few consonant changes.

With respect to the vowel changes, we find the most conservative varieties in the far north in Group 1 (Coatzospam and Cuauhtemoc) and again in Yucuhiti and Itundujia (both Group 6). All of the most innovative varieties are found in Group 7, again around Ixpantepec Nieves.

The northern varieties of Group 74 and Linkage 75 are much more conservative though. Unlike with consonantal changes, the coastal varieties are not particularly conservative when it comes to vowels, especially not those in the eastern part of the group. It is not surprising that these overall counts do not completely line up with the observations from Bradley & Josserand (1982), since their account is based on a handful of changes identified as important. The general tendencies we observe, however, do point in similar directions. We also find that overall, the varieties of the Baja region are more innovative and the coastal varieties and those in the far north-east are the most conservative.

Understanding which changes could potentially be used for improving subgrouping is of great importance for further unraveling the linguistic and cultural history of the Mixtec language family. It is equally instructive to better understand which changes are correlated with each other, which can indicate that they took place at the same time or that one change acted as a trigger for another. This in turn is a first step towards the broader goal of establishing a complete chronology of the sound changes identified in this study. The relative chronology of the sound changes is important because it can help shed light on the complex migration history of Mixtec people.

Establishing this detailed chronology lies outside the scope of the current study, but we provide a starting point by identifying the changes which correlate most with each other, as well as those which align best with subgroups and dialect areas. In the previous sections (3.4 and 3.5) when discussing the changes from each reconstructed Proto-Mixtec phoneme, we mentioned when a change aligned perfectly or very well with one or more subgroups form Josserand (1983) or Auderset et al. (submitted). We also mentioned other changes that are affected by the same environment or important with respect to the ordering of changes. We now summarize this information across the whole data set with a correlation coefficient, a cluster analysis, and a principal components analysis.

Since we are dealing with categorical and unordered variables, we calculate Cramér's V as a measure of correlation. Cramér's V is based on Pearson's chi-squared statistic and results in a value between 0 and 1. A value of 0 indicates that there is no association between the variables, while a value of 1 indicates that they are perfectly correlated. We calculate Cramér's V for each pair of sound change variables, as well as for each sound change variable with both the dialect areas from Josserand (1983) and the subgroups from Auderset et al. (submitted).

There are a handful of changes that are perfectly correlated with each other (Cramér's V = 1), all of which single out specific varieties or pairs of varieties: G03 (medial glottal stop insertion) and J14 (loss of initial *_j_ before _e_ are found only in Coatzospam. I08 (*_i_ to _ɨ_ before _a(ʔ)w_) appears only in Itundujia, while J10 (loss of *_j_ between _e_ and _i_ with glottal stop) is absent only in this variety. X03 (*_x_ to _j_ before _ɨ_) and X19 (*_x_ to x^j before e) are only present in Itundujia and Monteverde. W09 (loss of *_w_ between _a_ and glottal stop) and A06 (_a_ to _o_ in the context) only appear in Yutatio and Atenango. San Martin Duraznos shows a series of specific palatalizations (ND07, T10, X01) and one vowel change (U20) only present in that variety. Of the changes that are very highly correlated with each other (Cramér's V >0.9), many concern the same Proto-Mixtec sound in very similar environments (I05 and I06, O05 and O06, IB01 and IB02 and IB03). These changes are potential candidates for being merged together, should more data reveal a perfect correlation. A table with all correlations can be found in the supplementary materials.

In Figure 3.30 each data point represents the correlation coefficient of one sound change variable with one of the two classifications. The values are spread from about 0.2 to 1 and there is substantial overlap between the two classifications. Overall, the correlations are slightly stronger with the subgroups from Auderset et al. (submitted) than with the dialect areas (Josserand 1983), also reflected in the higher median. There are no generally agreed upon cut-offs for interpreting Cramér's V as indicating a 'weak' or 'strong' correlation. We set 0.6 as the threshold above which we consider changes to be strongly correlated with the groupings. We consider values between 0.3-0.6 as moderately correlated and below 0.3 as

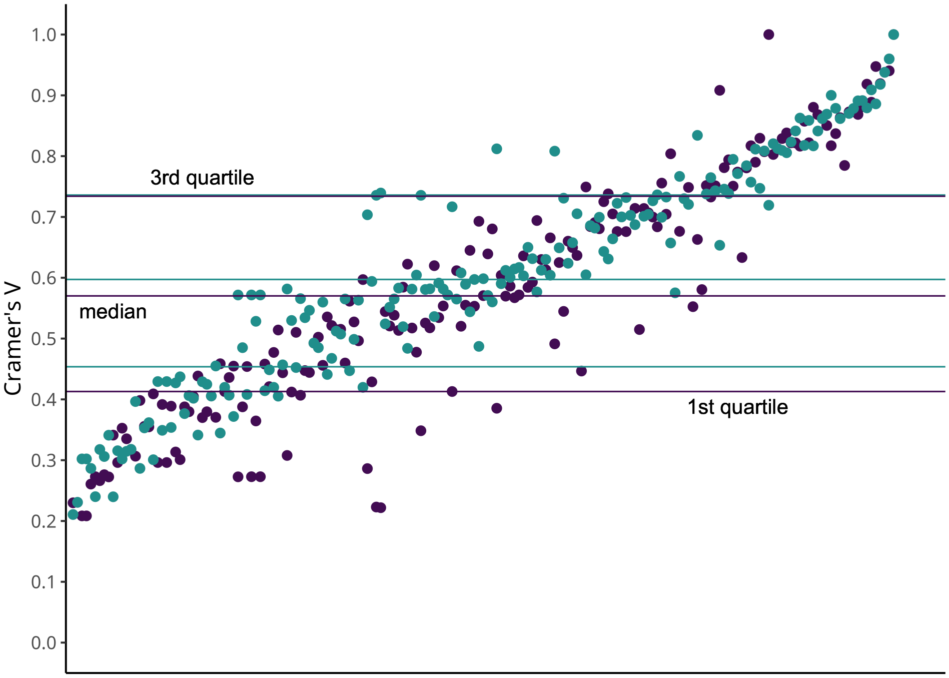


Figure 3.30: Association of each sound change variable with Josserand's (1983) dialect areas (purple) and Auderset et al.'s (submitted) subgroups (teal) with mean, median and first and third quartile

weakly or not correlated. Given these thresholds, 88 (45%) sound change variables show a strong correlation with Auderset et al.'s (submitted) subgroups and 83 (43%) with Josserand's (1983) dialect areas, with 72 of the changes overlapping (Table 3.15).

It is difficult to judge whether the division observed in Table 3.15 is expected or not, since we are not aware of studies applying the same methodology to other language families. From the viewpoint of characterizing Mixtec as a dialect area, where languages have remained in contact over time (cf. Section 4.1) it is surprising that only 15% or less of the changes are only weakly or not correlated with Josserand's (1983) dialect areas, but even more so with the subgroups established based on cognacy with computational methods Auderset et al. (submitted). This comparison indicates that the agreement between classifications of Mixtec languages based on various methods and data sources is quite high.

To further explore and visualize which languages are similar to each other based on the

Table 3.15: Number and percentage of sound changes by category of Cramér's V

Cramér's V	Josserand		Auderset et al.		Overlap
strongly correlated	83%	(43)	88%	(45)	72%
moderately correlated	83%	(43)	85%	(44)	64%
weakly/not correlated	29%	(14)	22%	(11)	14%

coded sound change variables, we apply clustering to the data. Cluster analysis refers to the task of grouping a set of data points into clusters (groups, in other words) such that the data points in each cluster are more similar to each other than to those in other clusters. There are many different algorithms and measures of similarity, depending on the type of data and question. Cluster analysis was first introduced in anthropology by Driver & Kroeber (1932) to find similarities in cultural practices. Here we apply agglomerative hierarchical clustering, also referred to as agglomerative nesting. This algorithm starts with each data point representing a cluster on its own. Clusters are merged at each step with the nearest cluster, until all data points form one cluster. The distance between clusters can be specified in different ways. We implement the 'complete linkage' method, in which the distance between two clusters reflects the distance between the two elements that are furthest away from each other. The result is a dendrogram, provided in Figure 3.31 that shows the hierarchical clusters obtained with coloring by Auderset et al.'s (submitted) groups. The same dendrogram colored by Josserand's (1983) dialect areas is provided in the supplementary materials X. We first note that the groups are recovered quite well, which is not surprising given the generally high correlations of the distribution of sound change variables with classifications obtained otherwise.

The first partition neatly separates Groups 1-6 from Group 7, or in other more commonly used terminology, it neatly separates the Alta varieties from the those in the Baja. The second partition as well as the lower partitions do not separate pre-established groups to the same extent, but degree of agreement between the clusters and the groups established based on

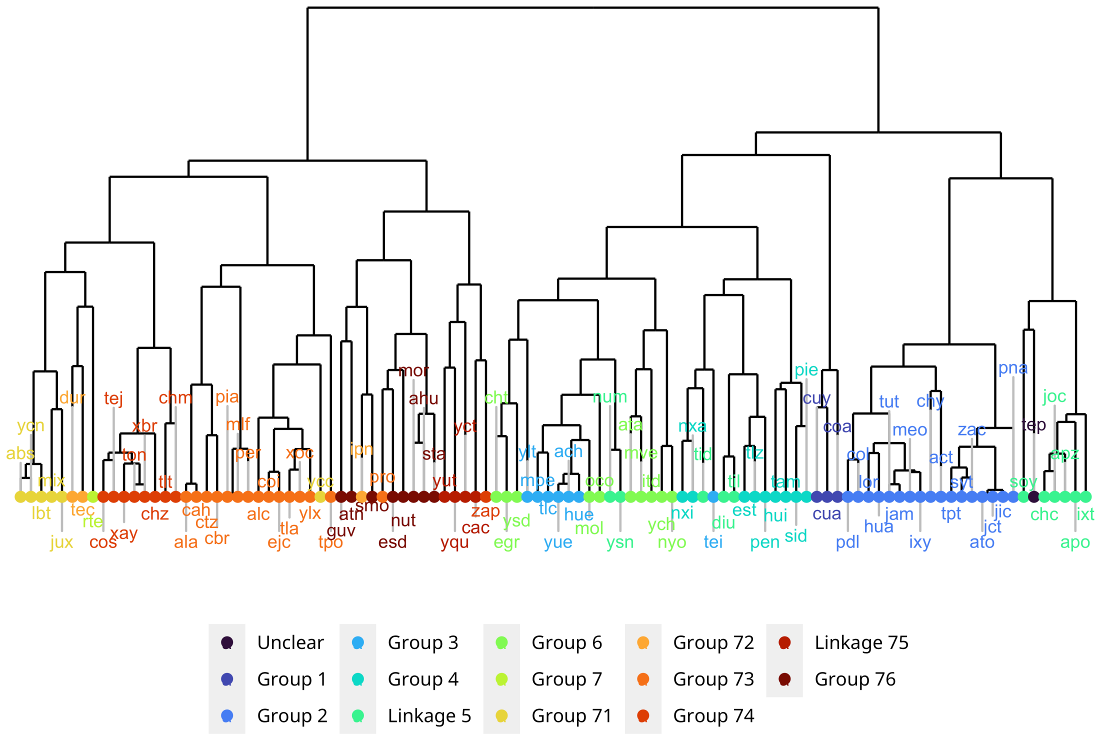


Figure 3.31: Dendrogram of agglomerative nesting with tips colored by Auderset et al.'s (submitted) groups

cognacy data with phylogenetics is still quite high.

As another way of exploring the data, we perform a principal component analysis (PCA) with the *pcaMethods* package (Stacklies et al. 2007) in the statistical programming language R (R Core Team 2022), which simultaneously imputes missing data. We apply the bpca method, which implements a Bayesian PCA missing value estimator. PCA is a method for data exploration, which is useful for summarizing information of large data sets such that this information can be visualized and analyzed more easily. The important information extracted from multivariate data is expressed through summary indices, so-called principle components. PCA is a non-parametric method and it can deal with multi-collinearity and missing values, as we have in our data set. Often, PCA is used for variable selection, but it is also an effective tool for discovering and visualizing hidden structure in data.

Figure 3.32 shows the result of the PCA. We ran the model with two components and the cumulative R2 for the first and second component is at 0.18. The x- and y-axis represent the first and second principal component, respectively. We flipped both axes, so that the distribution of the varieties lines up more closely with their geographical distribution. Varieties closer to each other in the plot are more similar to each other based on the sound change data, those far away from each are very different from each other. Additional figures that show different parts of the plot to render all labels legible can be found in the supplementary materials, as well as the same result colored by Josserand's (1983) dialect areas. Overall, the PCA results not only approximate the geographic distribution of the varieties, but the varieties also cluster to some degree along the lines of the subgroups from Auderset et al. (submitted). Our interpretation of this is that the sound changes as we coded them – that is on a very fine-grained level – generally align well with the cognacy data. As mentioned above for the clustering results, this is somewhat unexpected given the assumption that dialect areas are (implicitly) characterized by mismatches not only between various sound changes but also of sound changes with cognacy overlaps.

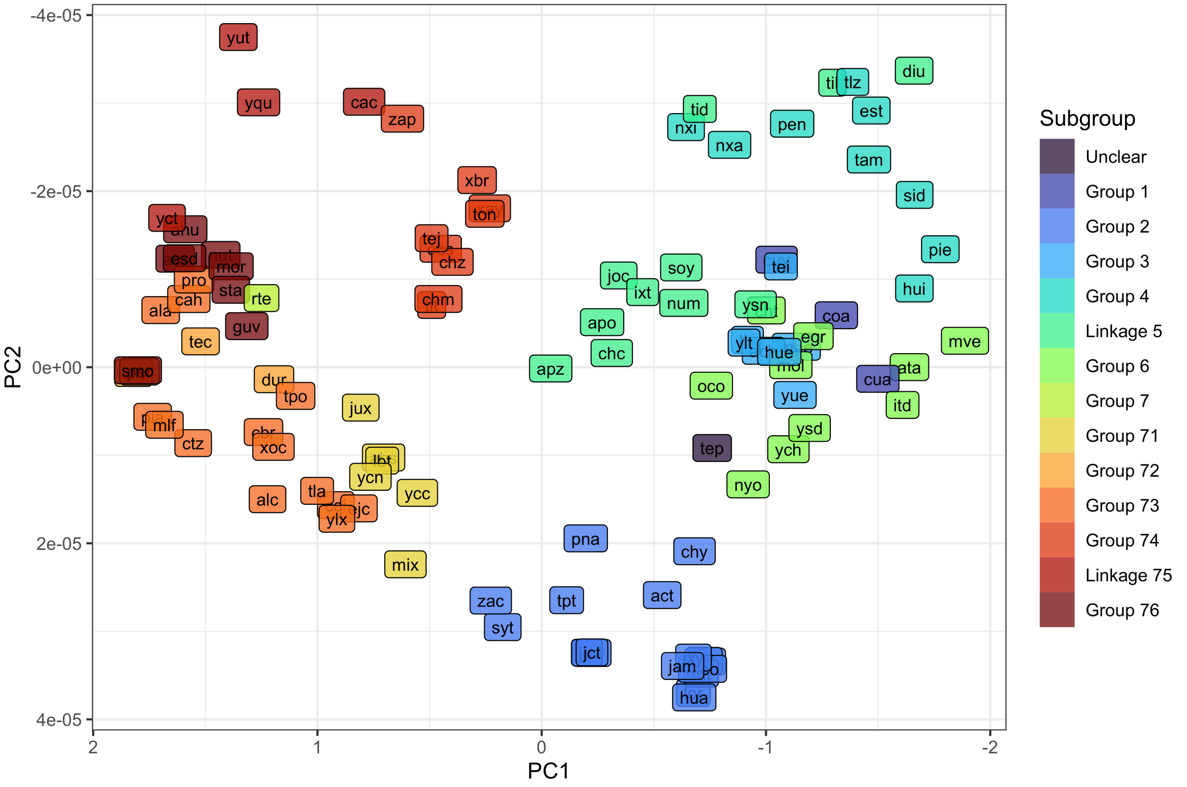


Figure 3.32: Score plot of the PCA with varieties colored by subgroups (Auderset et al. submitted)

In the upper right quadrant, we find the varieties of Group 4 and three varieties from Linkage 5. In Josserand's classification, all these varieties belong the same group (Eastern Alta). Remember that we use the term linkage to refer to varieties that were not firmly placed in one group or another in the consensus tree, in other words, our model did not exclude those varieties from being part of Group 4, the data just had conflicting signal. Given the result of the cluster model presented above and the PCA here, we conclude that Josserand's (1983) proposal is correct and these three varieties should be classified as Group 4. The rest of Linkage 5 is found in a relatively tight cluster in the center right area. These varieties correspond to Group 51 in Auderset et al. (submitted) and the Northeastern Alta area in Josserand (1983). Of the two remaining varieties of the linkage, Ñumi and Yosonama, are classified as Western Alta (roughly Groups 3 and 6) and especially the latter does appear closest to Chalcatongo, a variety from Group 6. The former seems closer to Group 51 based on the sound changes and further research is needed to clarify the best classification for those two varieties. Adjacent to the linkage, Group 3 varieties form a relatively tight cluster surrounded by varieties of Group 6. Both Teita and Yucuañe are located further away from the other varieties, the former closer to the Group 1 variety Cuyamecalco, the latter closer to the Group 6 variety Yosondua. The varieties of Group 6 are somewhat scattered, but like in the consensus tree Yucuhiti and Nuyoo are close together, while Ocotepec is relatively distant from all other varieties in the group. As mentioned above, Groups 3 and 6 form Josserand's Western Alta dialect area. Given the results of the PCA and the low posteriors connecting these groups in the consensus tree, it is possible that they form a larger group together. In the same area of the graph, we also find the three varieties from Group 1. Based on the previous studies, we expected Cuymecalco and Cuauhtemoc to appear close to each other and perhaps a bit further away from Coatzospam, but instead we find all three varieties relatively spread out. Further research should investigate whether this reflects different migration histories. What is also interesting is the position of the colonial variety of Teposcolula. Josserand (1983) classified it as Eastern Alta (our Group 4), while it was not conclusively grouped with any other varieties in the phylogenetic study. Here we find it closest to Nuyoo and Yucuhiti of Group 6.

In the lower right quadrant, we find the varieties of Group 2, which corresponds to the Coast dialect area. It is relatively well separated from all other varieties. On the upper edge, we find Chayuco as closest to Teposcolula and Nuyoo. On the left edge, we find Zacatepec relatively close to Mixtpec from Group 71. The latter is not surprising, since earlier studies proposed the Mixtepec area as the origin of the migration to the coast (Bradley & Josserand 1982:297-298). However, based on the sound change data, the languages are equally close if not closer to Teposcolula and some varieties of Group 6. This could point to multiple migrations rather than just one.

In the left half of the plot, we find all the varieties of Group 7. Again, we see the split between the Baja and Alta varieties replicated, which we already found in the clustering model. In the lower left quadrant, we find the varieties of Group 71, most of Group 73 and Linkage 72. In the upper left quadrant, we find the varieties of Groups 76 and 74 as well as Linkage 75. This broadly matches the geographic distribution of the varieties, especially in terms of latitude. The group that is mostly tightly clustered together and set off from the others is Group 74, which corresponds to Josserand's Northern Alta dialect area. This corresponds well to the relative geographic distance of these varieties, which are partly located in the state of Puebla. In both the cluster model and here in the PCA, Zapotitlan is separated from the rest of this group and placed closer to the varieties of Linkage 75, here especially Cacaloxtepec. In the consensus tree, the nodes connecting these two varieties to others have very low posteriors and they are relatively close to each other in the tree. Further research should investigate whether Zapotitlan should be reclassified together with Cacaloxtepec. Other varieties of Linkage 75, which corresponds to Josserand's Tezoatlan dialect area, are not spread out from each other but are quite far from all other varieties. This difference in terms of sound changes could help explain why they are difficult to classify. The varieties of the other groups within

Group 7 are closely intertwined. There is a relatively tight cluster of varieties of Group 74, but it is intertwined with varieties from Group 73 and Linkage 72. To sum up, the developments of the fine-grained sound changes as identified and analyzed in this study in general line up quite well with the classification obtained through cognacy data with Bayesian phylogenetics (Auderset et al. submitted).

Limitations: Cuicatec (for lack of sufficient data) and Triqui (for lack of a reconstruction that appropriately incorporates the divergent Itunyoso variety, cf. Matsukawa 2005) are excluded from the present study, leaving an updated Proto-Mixtecan reconstruction to future research.

3.7 Conclusion

In this study, we identified 206 segmental sound changes and their conditioning environments across 104 Mixtec languages. In this process, we were able to confirm many previous reconstructed Proto-Mixtec forms, refine others and add new ones. We provided a discussion of the important developments of each Proto-Mixtec sound (excluding tone) and provide the full data to address future research questions. While many open questions remain, the patterns and distributions of sound changes provided here lay the groundwork for an in-depth and updated understanding of the linguistic history of the Mixtec languages. First, we find support for the idea that what is commonly referred to as the Alta and Baja regions is not just a geographic distinction, but linguistically meaningful. The Alta region covers all groups except Group 2 (Coast) and Group 7 (Baja). This division is also suggested by ”a phylogenetic analysis in Auderset et al. (submitted) and is supported by many sound changes that only affect Group 7 or conversely, only Groups 1 to 6. It is also reproduced in the cluster analysis and visible in the PCA. This means that we now have multiple strands of evidence for a linguistic division between the Alta and the Baja varieties. It is important note that this division

– if it can be conclusively established – must be old, which is to say that the varieties within each region still exhibit a great deal of variation and are not generally mutually intelligible. The early migration to the far north-east of Group 1 is also well supported by our data. There are a few sound changes unique to the three varieties of this group, which is otherwise very conservative. This lends to an interpretation that Group 1 split off from the rest of Mixtec speakers early on and has remained relatively isolated since then. The overall conservatism of the coastal varieties similarly points to a relatively early migration. There are some indications that this migration could have started from the Mixtepec varieties (Group 71), such as the affrication of Proto-Mixtec **x*, which is unique to Group 2 and Group 71 varieties, but they are not conclusive. The position of the colonial variety of Teposcolula remains unclear. This is not surprising given that it was spoken in the border area between the Baja and Alta, in a region where Mixtec is no longer spoken today and we thus lack contemporary data. It is thus quite possible that it does not belong to any of the subgroups that comprise the contemporary varieties. Given that it was used as a *lingua franca* in colonial times, it could well have acquired features of multiple other varieties. Within the large Group 7, we found multiple changes that clearly delineate Group 71 and Group 74 from the rest. In the case of the latter, the relative geographic distance of these varieties spoken in the north in Puebla or close to the Oaxaca-Puebla border explains why they are often more conservative than other varieties of Group 7. To sum up, the fine-grained sound changes analyzed in this study confirm that the comparative method applies to this language family just as any other. The distribution of the changes themselves reveal more tree-like and more wave-like parts of the family and generate a host of new questions for further investigation. But the sustained contact between the varieties does not invalidate historical work on this language family. Likely the situation found in the Mixteca is not as different from that of other, better studied language families.

The advent of computational and quantitative methods in historical linguistics has advanced the field, but also led to a (perceived) divide between those approaches and more traditional ones. We hope to have shown that the former does not preclude or compete with the latter. While our study in essence is qualitative, making use of the comparative method for identifying sound changes and conditioning environments, we also draw on computational data visualization and analysis techniques. These methods are invaluable when dealing – as we do here – with a large number of changes and languages, which are difficult to summarize and compare by hand. By applying methodology from autotypology and providing the data collected and analyzed, we hope to encourage more detailed studies of subgroups within Mixtec, but also studies on other Otomanguean branches. Thanks to recent and often collaborative documentation efforts in the Mixtec region and Mesoamerica more broadly, we have better and more descriptions of the languages there. We are thus in a good position to make significant advances with respect to the linguistic and cultural history of the Mixtec and other indigenous people.

3.8 Data Availability

The supplementary materials to this chapter can be found at https://osf.io/7ar39/?view_only=9567daca035a4434ba9e11d32ac0d510.

Chapter 4

Rates of change and phylogenetic signal in Mixtec tone

Disclaimer

A slightly modified version of this chapter was submitted to the Journal of Language Dynamics and Change as: Auderset, Sandra. Rates of change and phylogenetic signal in Mixtec tone

4.1 Introduction

Historical linguistics as a field has expanded considerably in the past decade both with respect to methodology and data. The adoption of computational and quantitative methods, such as automatic sequence alignments (List et al. 2018) and Bayesian phylogenetics (Greenhill et al. 2020), has led to more explicit and testable hypotheses and exchange with other fields dealing with the human past, such as archaeology and population genetics. Language documentation efforts all over the world have made it possible to diversify the data used to study language change. There is, however, a gap in the field when it comes to tone. Tonal phenomena are conspicuously absent from studies on language change despite the abundance of tonal languages worldwide (Yip 2002). Research on tonal correspondences (Dockum 2019), tone reconstruction (Dimmendaal 2011), and tone change (Janda & Joseph 2003) is still relatively rare and often confined to the emergence of tonal contrasts from segmental changes (Campbell 2021). The latter is reflected in a long tradition of work on tone correspondences in languages of Southeast Asia, which relies on established processes of tonogenesis (Gedney 1972, Joseph & Burling 2001, Ferlus 2004, Dockum 2019, among others). But for tone change apart from tonogenesis no such body of literature exists.

It has often been observed that tone realizations can vary drastically even in closely related languages (Cahill 2011, Beam de Azcona 2007, Morey 2005, Dürr 1987, among others), leading to the assumption that tones change faster than segments (Ratliff 2015). This in turn has led to tone being omitted in studies on the history of a language family (Campbell 2021). Further complications arise with respect to the comparison of tonal systems more broadly. Available typological studies are coarse-grained and based on highly analysis-dependent categories such as the number of tonal contrasts or wholesale classification of tonal systems as 'simple' vs. 'complex' (Lee 2022, DiCanio & Bennett 2020, Maddieson 2013). To sum up, there are many challenges with respect to research on the historical dimensions of tone and empirical studies on the rate of tone change and its contribution to subgrouping outside of tonogenesis are still lacking.

In this study, I take a first step at addressing this gap. I investigate tone change in Mixtec combining data analysis practices from typology and quantitative methods from historical linguistics. I compare phylogenetic signal and rates of change in tones versus segments to empirically address the claims about tone volatility mentioned above and assess whether tones indeed change faster than segments and show less phylogenetic signal. The Mixtec languages of southern Mexico, which form part of the Mixtecan family of the Otomanguean stock, provide an ideal testing ground for such an investigation. These languages are all tonal and tone carries a high functional load both in lexicon and grammar. Tones in Mixtec are old and must be reconstructed to proto-Mixtec and proto-Mixtecan and most probably as far back as proto-Otomanguean (Campbell 2017b, Rensch 1976). Furthermore, there are previous historical linguistics studies on Mixtec including reconstructions of the proto-Mixtec segmental inventory (Longacre 1957, Mak & Longacre 1960, Bradley & Josserand 1982), identification of segmental sound changes (the aforementioned and Josserand 1983, Auderset & Campbell in prep), but also tone reconstruction and tone change (Dürr 1987, Swanton & Mendoza Ruíz 2021). Due to recent intensification of documentation efforts in the Mixteca region and diaspora communities, we now have data for more varieties than the earlier sources were able to draw on and can extend or update previously collected data on a number of varieties. In addition, we have a recent, up-to-date family tree of Mixtecan based on Bayesian phylogenetic methods (Auderset et al. submitted), which provides considerable detail concerning the lower-level relationships of these languages.

As mentioned above, comparison of tones in general is challenging even on a synchronic level. To address these challenges, I draw on data organization and analysis methods developed for multivariate typology (Bickel 2010, Bickel et al. 2011, Bickel 2015). Multivariate typology works with systems of fine-grained variables which are created bottom-up from the language data and thus aim at capturing all the variation present in the data without imposing preconceived definitions (Bickel 2010). This approach is easily adaptable to diachronic data, such as sound changes, and has the added benefit of resulting in databases that can be re-used for other research questions. I thus establish sound changes, both tonal and segmental, in Mixtec and code them in interlinked databases. The process is described in more detail in section 4.2. To address the questions of whether tones change faster than segments and whether or not they can be used for subgrouping in an empirical way, I rely on computational methods established for exploring biological evolution but which are currently gaining traction in studies on linguistic change (Macklin-Cordes et al. 2021, Hübler 2022, Phillips & Bowern 2022, among others). Phylogenetic signal measured by the metric *D* is based on the sum of sister-clade differences across a family tree. It provides a measure for the tendency of related species (sisters in the tree) to be more similar to each other than to other species sampled randomly from the tree (Fritz & Purvis 2010, Münkemüller et al. 2012). In linguistic terms, phylogenetic signal reflects the tendency of closely related languages to be more similar to each other with respect to a given variable than to more distantly related languages in a given language family tree. Comparing phylogenetic signal across segmental and tonal changes thus provides us with an objective measure of how much these changes align with and identify language relationships within Mixtec. This in turn will help answer the question of whether tones carry phylogenetic signal, or in other words, whether tone changes can (and should) be used for subgrouping alongside segmental changes. To investigate evolutionary rates of change in tones and segments, I use a Hidden Markov Model. This model estimates transition rates between two observed characters based on a tree. The 'hidden' part of the term refers to the fact that in this type of model multiple processes can be invoked to describe the evolution of the observed characters (Beaulieu et al. 2013). Applied to linguistic evolution, this means that the model estimates the transition rates (rates of change, in other words) between the presence and absence of a variable across the language of the family tree. As opposed to other models, it allows the transition rate to vary across the tree, such that a variable can change quickly near the root of the tree and then slow down in some branches, for example. This suits changes in linguistic structures well, as those are not assumed to proceed at a constant, unchanging rate (Nettle 1999).

Given that the majority of the world's languages are tonal (Yip 2002) and historical linguistics is one of the primary ways of accessing the past of an ethnolinguistic group, understanding whether or not tone can contribute to that is crucial. Earlier studies on tone correspondences suggest that it does, but to what extent remains unclear. Assumptions on

Table 4.1: 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

the volatility and variability of tone suggest that it does not. If we can show that tone change proceeds at similar rates as segmental change and that tone changes carry just as much phylogenetic signal, this will make a strong case for no longer ignoring this important aspect of language in historical studies.

4.2 Data collection and analysis

The data used for this study constitutes a subset of an existing database of annotated cognate sets across 130 Mixtecan languages (Anonymous 2022). These cognate sets were established based on a word list of 209 concepts tailored to the Mesoamerican cultural area. Verbal forms were excluded because they require aspect-mood inflection, which is not well understood on a comparative Mixtec level, and is not always provided or reliably identifiable in glosses in the source material. The entire list is provided in the supplementary materials (and is identical to the one that is used in the previous study, see Auderset et al. submitted for more details). In a next step, the orthographic entries were converted to IPA and tone notation was standardized. I only discuss the standardization of the tone notation here, other details can be found in the supplementary materials 2 of Auderset et al. (submitted), Anonymous (2022).

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 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, cf. Table 4.1. 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. The source materials display a wide range of tone notations. Most descriptions of varieties with three tonemes denote them with diacritics, while those with more tonemes usually represent them with numbers. The mapping of a specific diacritic or number to toneme varies widely in the sources. The tone notations found for level tones in the sources and their standardization used in all the materials of this study are summarized in Table 4.2.

For the present study, I restricted the data to just those Mixtec varieties for which reliable information on tonal contrasts is available. Out of the 130 Mixtec varieties for which we were able to collect entries for the word list in the earlier study (Auderset et al. submitted), only 46 have tones marked on more than a few entries.[1] Of these 46, Metlatonoc and Molinos Mixtec had to be excluded because of low coverage (defined here as NA for more than half of the variables), and Diuxi and Abasolo del Valle due to difficulties in the interpretation of the tone values (see supplementary materials for details), resulting in a sample of 42 languages. These 42 languages belong to 6 of the 7 larger subgroups identified in the previous study (Auderset et al. submitted) and to 10 out of the 12 dialect areas identified by Josserand (1983).

[1] Note that I do not distinguish between languages and dialects, since there is no solid basis nor necessary or sufficient criteria to do so in Mixtec. I 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).

Table 4.2: Tone levels and standardized tone notation with an overview of notations found in source materials

Description	**5**	**4**	**3**	**2**	**1**
	high	mid-high	mid	mid-low	low
all three marked	acute (á)		macron (ā)		grave (à)
mid-low grave, low underbar	acute (á)		unmarked (a)	grave (à)	underbar (a̠)
low unmarked, mid macron	acute (á)		macron (ā)		unmarked (a)
low unmarked, no mid	acute (á)				unmarked (a)
mid unmarked, low macron	acute (á)		unmarked (a)		macron (ā)
no mid, low macron	acute (á)				macron (ā)
mid unmarked, low grave	acute (á)		unmarked (a)		grave (à)
no mid, low grave	acute (á)				grave (à)
mid unmarked, low underbar	acute (á)		unmarked (a)		underbar (a̠)
Chao	5	4	3	2	1
Chao with 4	4		3	2	1
Chao with 3	3		2		1
inverse Chao	1	2	3	4	5
inverse Chao with 4	1	2	3		4
inverse Chao with 3	1		2		3

Due to lack of documentation, I was not able to include any varieties of Linkage 5 (roughly corresponding to Josserand's (1983) Tezoatlan area, but inlcuding a few more varieties). A list with all varieties, subgroup and dialect area affiliation, abbreviations and language codes is provided in the supplementary materials. Figure 4.1 provides a geographic overview of our sample with subgroup affiliation and Figure 4.2 shows the maximum clade credibility tree from Auderset et al. (submitted) pruned to just the languages under discussion here, illustrating the genealogical relationships of the varieties in more detail.

Reconstructed segmental proto-forms for these cognates were also available from a previous study (Auderset & Campbell in prep). Those were created using the comparative method and based on already established reconstructions provided in Josserand (1983) and Dürr (1987). Each reconstruction was then carefully rechecked and reconsidered in light of the newly available primary data and – if needed – updated to match the current data set and knowledge of the sound changes. I also added new lexical reconstructions not covered by previous sources.

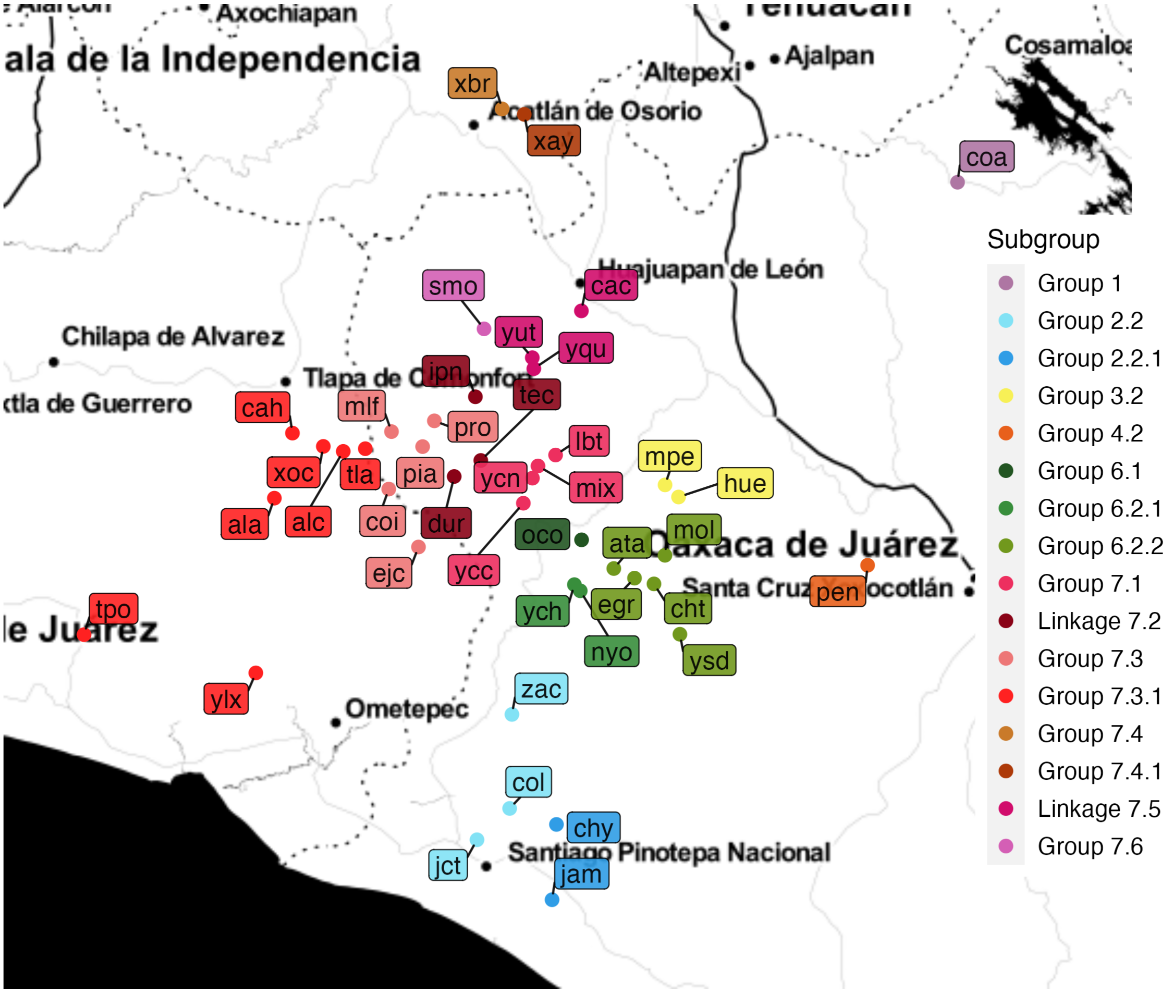


Figure 4.1: Languages sampled with subgroup affiliation (according to Auderset et al. submitted)

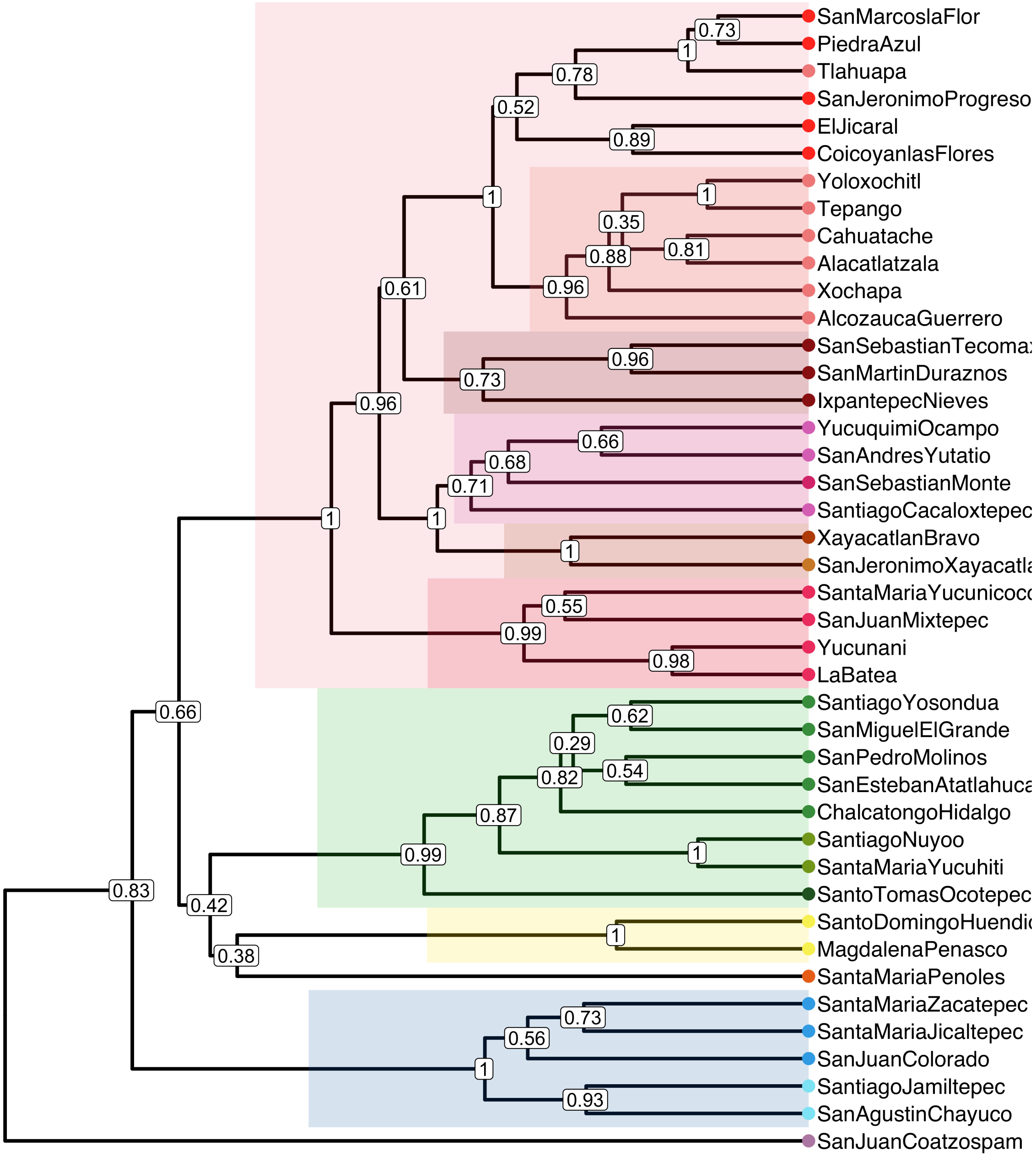


Figure 4.2: Maxiumum clade credibility tree of the sample colored by groups (see map legend for details)

For details regarding these updates, see Auderset & Campbell in prep. It was not always possible to arrive at a single, most probable proto-form with the material available. This most often concerns reflexes of **u* and **o*, which are difficult to disentangle, especially when nasalized. In such cases, I provide an alternative proto-form and take the uncertainty regarding the vowel into account when coding the sound change variables. All the reconstructions, alternative proto-forms, and earlier proposals can be found in the supplementary materials.

Mixtec languages have a strong tendency for lexical morphemes to be bimoraic, and this applies to reconstructed lexical proto-Mixtec morphemes as well. The bimoraic unit is called the 'couplet' in Mixtec literature (Pike 1948), a terminology I follow. Each vowel counts as one mora and long vowels consist of two morae. There are no final consonants apart from the glottal stop, which is retained only in two varieties, and consonant clusters are very restricted or absent in all modern Mixtec varieties, which means that most couplets have a syllable structure of CVCV, CVV, VCV, or VV. Lexical entries can also have more than two moras (just not less). Most commonly these longer forms have three moras and historically consist of a one-mora (CV or V) 'prefix' added to a bimoraic couplet. This pattern is particularly common in animal names, which often need to be reconstructed with the animate classifier **ti*. I reconstruct both bimoraic and trimoraic forms, segmenting the latter into the 'prefixal' element and the couplet. I removed duplicates and all entries which are not tone marked, as well as lexemes that only appear in one language in the data set. This results in a total of 5,255 data points spread over 262 cognate sets. Coverage per variety is between 34 to 197 cognate sets, with a mean of 122 and a median of 125 sets.

4.2.1 Segmental correspondences and variables

In order to compare segmental and tonal change across the varieties of our data set, I identified segmental correspondences and derived segmental change variables based on the reconstructed proto-forms and the cognate sets. Segmental correspondences, changes, and reconstructions are better understood than tonal ones and have been studied in more detail (Bradley & Josserand 1982, Josserand 1983, Auderset & Campbell in prep). The databases containing segmental proto-forms, cognate sets, and segmental change variables constitute a subset of those created for a previous study on the distribution of those changes (Auderset & Campbell in prep). I summarize the methodology only briefly, as all the details are laid out in Auderset & Campbell in prep.

I create multiple, interlinked databases following AUTOTYP principles such as modularity, autotypology, separation of definition and data files, and late aggregation (Witzlack-Makarevich et al. 2022). The AUTOTYP research project was developed to address problems that arose 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 traditional usage that often fail to capture a phenomenon across a large and diverse sample of languages. I adapt these principles and guidelines to ensure that the data set is expandable in the future, re-usable for other research questions and created in a bottom-up fashion that allows for maximal accuracy and transparency.

Based on the established cognates sets, I identified regular sound correspondences linked with a proto-Mixtec reconstructed sound. From these correspondences, I derived binary segmental change variables and coded these variables for each variety of the sample. I illustrate the process and coding with a small example; in Table 4.3, I provide four cognate sets with Proto-Mixtec reconstructions and reflexes from three varieties. We can see changes in the final vowel of DISEASE and loss of the glide **j* in the same set, as well as in GRIDDLE. Set TWO shows loss of the glide **w* and a different set of final vowel changes. Finally, GRIDDLE and TORTILLA illustrate different reflexes of Proto-Mixtec fricatives and effects of palatalizations. Table 4.4 summarizes the sound changes and coding identified for the sample for three contemporary varieties.

Table 4.3: Four example cognate sets across three Mixtec varieties

Mixtec Variety	DISEASE	TWO	GRIDDLE	TORTILLA
Proto-Mixtec	*k^{w} e^{B} ʔ j i^{B}	*u^{B} w i^{B}	*x i^{B} j o^{B} ʔ	*s i^{B} t a^{B} ʔ
Santa María Peñoles	k^{w} e^{5} ʔ e^{5}	u^{5} u^{5}	ʃ i^{5} o^{5}	ð i^{5} t a^{5}
San Esteban Atatlahuca	k^{w} e^{3} ʔ j i^{1}	u^{3} u^{1}	x i^{3} ʒ o^{1}	s t a^{3} a^{1}
Xochapa	k^{w} e^{1} ʔ e^{1}	u^{1} β i^{1}	ʃ i^{1} j o^{5}	ʃ i^{1} t a^{5}

Table 4.4: Example segmental variables derived from the cognate sets in Table 4.3

Mixtec Variety	*j >ø/_i	j >ø/i_o	j >ʒ/_o	*w >ø/_i	*x >ʃ/_i
Santa María Peñoles	yes	yes	no	yes	yes
San Esteban Atatlahuca	no	no	yes	yes	no
Xochapa	yes	no	no	no	yes

	*i >e/j_	*i >u/w_	*i >ø/s_t	*s >ʃ/_i	*s >ð
Santa María Peñoles	yes	yes	no	no	yes
San Esteban Atatlahuca	no	yes	yes	no	no
Xochapa	yes	no	no	yes	no

Of the segmental change variables and coding established in this way, I took a subset of the data that includes only the 42 varieties used in this study. I exclude variables that have data for less than half of the varieties and those which show the same value across all languages of the sample (that is, they are either present or absent in all sampled languages). The latter is necessary because variables that have the same value across all languages do not contribute information about subgrouping, splits, and rates of change. This results in a total of 184 segmental change variables, of which 81 pertain to vowel correspondences and 103 to consonantal correspondences. The complete data set is provided in the supplementary materials and a more detailed explanation of the coding method and process is provided in (Auderset & Campbell in prep).

4.2.2 Establishing tone correspondences and deriving tonal change variables

There are many obstacles to comparing and reconstructing tones. First, there is often little reliable data, as it is still common in descriptive materials of tonal languages to either ignore tone completely or only note it in a few cases (Campbell 2021, Beam de Azcona 2007), e.g. for the disambiguation of segmentally identical forms. In materials that do mark tone consistently, the phonetic realization can be unclear due to variation in tone notation and analysis (see also above in Section 4.2). Descriptions of tone in Mixtec languages are often confined to a list of the inventory and how each identified tone is marked in the orthography. Only rarely are minimal pairs provided, and discussions of allotones or distribution of tones are even rarer, especially in older materials. Fortunately, this situation seems to be shifting, with some recent sources on Mixtec varieties providing rich discussions of tone, see McKendry 2013, Mendoza Ruíz 2016, Vázquez 2017, Peters 2018.

Furthermore, there is little cross-linguistic research to draw on: We simply do not know which tone changes are common or even possible. In this regard, we have not significantly progressed from the situation 50 years ago, summarized by Ballard (1969:p.105;fn.10) as follows: “The two major obstacles to tone reconstruction to date have been a lack of knowledge of the phonetic nature of tones and a lack of experience in doing such reconstructions.” The latter is certainly still true today. The former, I believe, does not constitute an obstacle to the reconstruction of tone *per se*. When comparing tones in a historical perspective, we need to focus on contrasts (tonemes) rather than their phonetic realization, just as we reconstruct segments based on phonemes and not phones. Furthermore, we cannot know the exact phonetic realization of reconstructed segments either, even though the greater standardization through the use of IPA symbols can lead to the impression that we do. That proto-segments are also abstract notational conventions is most obvious in cases like the laryngeals of proto-

Indo-European, numbered from one to three due to uncertainty regarding their articulatory and acoustic properties (Beekes 1989). This is not to say that there are no proposals as to what phonemes the three laryngeals correspond to, but rather that this is a separate question which has not impeded reconstruction or the analysis of sound changes in this language family. An example from within Mixtec concerns the phonetic realization of Proto-Mixtec **x*: according to Mak & Longacre (1960:33), the sound that Longacre (1957) and others reconstruct as the proto-Mixtecan velar fricative **x* unconditionally lenited to **h* in proto-Mixtec. This is plausible as there are contemporary varieties which retain *x*, but note that this phoneme has an allophone *h* or that the pronunciation is closer to [h] generally (e.g. in Hollenbach 2013:11 and Macaulay 1996:20). Despite this open question, we can still establish the sound correspondences and reflexes of Proto-Mixtec **x*.

Before discussing the methodology of establishing tone change variables, I cover the basics of Mixtec tone systems and summarize two earlier studies of Mixtec tone reconstruction. The tone-bearing unit (TBU) in Mixtec is the mora, which means that each ‘couplet’ has two tones. These two tones can be level or contour tones, depending on the variety. In a historical perspective, there is a further distinction of ‘basic’ vs. ‘modified’ couplets already established in Longacre 1957. ‘Basic’ couplets are reflexes that do not contain ‘modified’ tones. These 6 sets – *high-high(glottal), *low-low(glottal), *high-low, *low-high – are said to be more regular than the modified ones (Dürr 1987:21). Dürr (1987) established tone correspondences across 17 Mixtec languages based on 110 cognate sets. He reconstructed two tones to proto-Mixtec (*high and *low) from which the contemporary tone reflexes are derived. The loss of the final glottal stop in most varieties then led to a third tone, often realized as a floating tone or a high tone (see Pankratz & Pike 1967), but in other varieties as a low tone (Campbell & Reyes Basurto forthcoming). The cognate sets with such a third tone correspond to the ‘modified’ couplets mentioned above. Dürr (1987) focused on the reconstruction of the basic sets. The tone reflexes of the 17 languages he analyzed fall into two groups: Area A, which covers the majority of languages, and Area B, containing only Peñoles, Diuxi, Jicaltepec, and Coatzospam Mixtec, with the latter showing divergent patterns altogether. Area A generally has a mid tone reflex for *high and a low tone reflex for *low, while area B exhibits low tone for *high and high tone for *low. The third (modified) tone is said to result in a high tone in area A, but has a lowering effect in area B (Dürr 1987:35). Dürr's (1987) study is explicitly preliminary and he pointed to several remaining issues that should be addressed with more data.

More recently, Swanton & Mendoza Ruíz (2021) take Dürr's (1987) reconstruction as the basis for investigating tones in Alcozauca Mixtec in a diachronic perspective. As opposed to most varieties included in Dürr's (1987) study, Alcozauca Mixtec has four contrastive tone levels, not three, but shows no tone alternations across word boundaries; in other words, there are no floating tones or tonal processes. The study also includes data from three other varieties: Santa Maria Zacatepec, Chalcatongo de Hidalgo, and Yucuquimi de Ocampo. With respect to the 'basic' sets, they confirm Dürr's (1987) correspondences and reconstructions. They are particularly interested in the 'modified sets', which the previous study did not systematically deal with. They add 10 modified tone groups, which they analyze as having consisted of a preposed glottal stop and/or tone followed by one of the basic sets. As mentioned above, the tone-bearing unit is the mora and most Mixtec lexemes are bimoraic, which means that generally a lexeme will have two tonemes. These combinations of two tones on a couplet are referred to as tone sets or tone melodies. I summarize all proposed tone reconstructions in Table 4.6. With respect to the basic sets, Dürr (1987:23) pointed out that the glottal stop "seems to be restricted to tonemic couplets with identical tones." This is maintained in Swanton & Mendoza Ruíz (2021). The modified sets are established based on the idea that tonal reflexes were not only affected by following glottal stops, but also by preceding glottal stops or tones. For example, some 'modified' tone melodies are the same as some derived melodies in verbal inflection, where the presence of the triggering tone can be observed on aspect-

Table 4.5: Overview of the six basic tone reconstructions proposed by Dürr (1987) and the ten modified tone reconstructions proposed by Swanton & Mendoza Ruíz (2021)

Basic		Modified	
1	*HH	7	*(ʔ1)-HH
		8	*(ʔ2)-HH
2	*HHʔ	9	*(ʔ)-HHʔ
3	*LL	10	*(ʔ)-LL (rare)
		11	*ʔ-LL (rare)
4	*LLʔ	12	*ʔ-LLʔ
5	*HL	13	*ʔ-HL
		14	*L-HL
		15	*(L)-HL
6	*LH	16	*Lʔ-LH

mood prefixes for some verbs. Other modified tone melodies can be explained by analyzing forms as earlier but now reduced compounds whose initial stem – and its original tone and presence or absence of glottal stop – is identifiable or even attested in colonial era sources (e.g. DOOR, from 'mouth of house', a widespread Mesoamerican calque, see Smith-Stark 1994).

One of the open questions concerns the two tone levels that should be reconstructed to proto-Mixtec. Longacre (1957), who worked on proto-Mixtecan including tone before Dürr (1987), reconstructed the two tones as mid and low, rather than high and low. The problem is best summarized with the help of Figure 4.3: the reflexes are basically opposite in area A and B, which means that whichever one chooses as the proto-form, one then has to explain the inversion in the other area. Unfortunately, we are not in a much better position to assess the issue today, since good tonal descriptions are still largely lacking for varieties within Dürr's area B. I thus adopt Dürr's (1987) tone levels of high and low in our study. As mentioned above, the fact that these reconstructed proto-Mixtec tones might not have been phonetically high and low does not preclude us from analyzing the tonal correspondences and deriving tone change variables, just as we do with segments.

Based on the cognate sets which were previously assigned a tonal reconstruction by Dürr

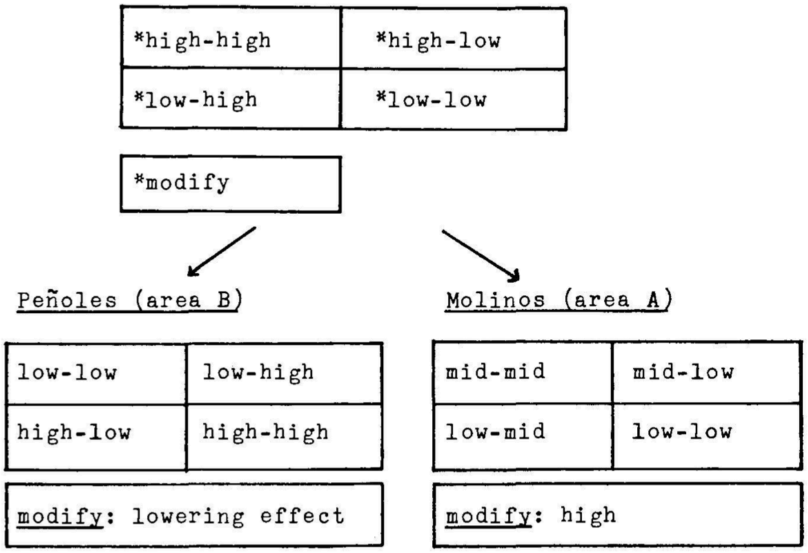


Figure 4.3: Proto-Mixtec tonemic couplets and reflexes in area A and B (reproduced from Dürr 1987:35)

(1987) and Swanton & Mendoza Ruíz (2021), I established tone correspondences across the 42 languages of the sample. I started with the basic sets, because there is ample data for each of those sets. The correspondences established in this way allowed me to identify other cognate sets belonging to the same tone group. I added 7 cognates sets for *HH, 17 for *HHʔ, 11 for *LL, 21 for *LLʔ, 7 for *HL, and 4 to *LH, which in turn helped refine the correspondences. I then reviewed the new sets proposed by Swanton & Mendoza Ruíz (2021) in light of the data set used in this study. The tone groups I identified in the modified sets overlap but do not fully agree with those from Swanton & Mendoza Ruíz (2021). This is hardly surprising, given that their reconstructions are based on four languages, while this study takes into account 42 languages. I refrain from positing specific tones or glottal stop as having preceded and conditioned the modified sets. Doing so would be very speculative, given that we do not understand floating tones and tone sandhi phenomena in modern Mixtec languages on a comparative level, let alone the history of these phenomena (see Beam de Azcona 2007 for a similar finding in a Zapotec subgroup). However, it is often still possible to posit which

Table 4.6: Schematic overview of reconstructed tone sets with labels

Basic		Modified	
A1	*HH	A3	*?.HH
		A4	*?.HH
A2	*HHʔ	A5	*?.HHʔ
		A6	*?.HHʔ
B1	*LL	(B3	*?.LL)
B2	*LLʔ	B5	*?.LLʔ
		B4	*?.LL?
C1	*HL	C2	*?.HL
		C3	*?.HL
		C4	*?.HL
D1	*LH	D2	*?.LH
		D3	*?.LH

basic couplet a modified set is based on. To still be able to distinguish between sets and environments (even if only on an abstract level), I labeled them with a capital letter and number combination, summarized in Table 4.6. It is possible that some of these 18 groups could be collapsed in the future if more data becomes available and we are better able to identify irregular tone reflexes and inconsistencies in source materials. More research might also show that some of the modified sets are actually based on a different basic set than I propose here, but note that this would not impact the correspondences as such.

Based on the correspondences found in the cognate sets and the 18 tone groups, I identified sound change variables in a bottom-up fashion.[2] The process is much the same as with segmental sound changes. For each unique reflex of a given Proto-Mixtec tone in a tone group, which represents a change, I create a variable and code the presence or absence of said change in each language of the sample. I illustrate this with a small example: Table 4.7 shows the reflexes of Proto-Mixtec *low tone in sets B1, B2, and B5 across four languages. Since I follow the reconstruction of tone values proposed in Dürr (1987), low tone reflexes are not coded as

[2] I did not code variables for tone group B3 since I have only cognate set in this group with few entries.

Table 4.7: Example tone correspondence sets

Mixtec Variety	B1 *L.L	B2 *L.Lʔ	B5 *ʔ.L.Lʔ
Santa Maria Peñoles	5.5	5.5	51.5
San Andres Yutatio	3.1	3.1	5.1
Yoloxochitl	1.1	1.5	5.25
Ixpantepec Nieves	1.1	1.1	5.5

Table 4.8: Example tone variables derived from the correspondences in Table 4.7

VariableID	PM	Reflex	Env.	SM Peñoles	SA Yutatio	Yoloxochitl	Ixpantepec N
Tone18	*L	3	#_	no	yes	no	no
Tone20	*L	5		yes	no	no	no
Tone24	*L	5	_ʔ#	yes	no	yes	no
Tone31	*L	5	(B5)#_	no	yes	yes	yes
Tone32	*L	5	(B5)_#	yes	no	no	yes
Tone33	*L	51	(B5)#_	yes	no	no	no
Tone34	*L	25	(B5)_#	no	no	yes	no

a change but viewed as a retention in these sets. In set B1, Yoloxochitl and Ixpantepec Nieves thus show no change, while Yutatio shows a change to mid tone in the first mora (Tone18) and Peñoles to high in both morae (Tone20). In set B2, most varieties show the same reflexes as in set B1, but in Yoloxochitl I find a high tone in the second mora, conditioned by the final glottal stop (Tone24). In set B5, where I do not know the conditioning environment before the low tone couplet, all varieties but Peñoles have a high tone in the first mora (Tone31), Peñoles and Ixpantepec also in the second mora (Tone32). Yoloxochitl has a contour tone on the second mora in this set (Tone34). Thus I identify 7 tone change variables based on this example set, which are summarized with the coding for each variety in Table 4.8. I apply the same methodology to all 42 varieties across the 17 correspondence sets. This results in a total of 61 tone change variables, of which 27 pertain to 'basic' tone sets and 34 to 'modified' ones. The full database is available as the supplementary materials.

4.2.3 Overview of basic tone correspondences

Before describing the methods and results in more detail, I present a qualitative overview of tone correspondences found in the data set. With respect to the basic couplets, the correspondences in this data set align quite well with Dürr's (1987) findings. Figure 4.4 summarizes the default reflexes of Proto-Mixtec *high and Figure 4.5 of Proto-Mixtec *low. While the reflexes do not neatly fall into two groups for either of the proto-Mixtec tones, the tendencies observed in Dürr (1987) are recovered in our sample. All varieties except Peñoles, Coatzospam, Tepango, and Jicaltepec have a mid tone reflex for Proto-Mixtec *high, while the five mentioned varieties have a low tone reflex. These two sets of varieties largely mirror Dürr's area A vs. B. The varieties with a mid tone reflex are further divided into five smaller groups depending on the effect of the glottal stop or absence thereof. In a cluster of varieties in the south-eastern part of the Mixtec region, commonly referred to as the Mixteca Alta, the final glottal stop has no effect. This also seems to be the case in Cahuatache Mixtec in the Baja region, but this variety is documented based on historical documents and the interpretation of the tonal values is not straightforward (Dürr 1987:fn.16, p.58). In the northern area, we predominantly find varieties in which the glottal stop has a lowering effect, resulting in a low tone reflex. In the south-western area, the glottal stop has a raising effect, such that these varieties show a high tone reflex in that context. These two groups of varieties are somewhat intermeshed and neither geographically nor genealogically neatly separated. There are two smaller divisions in which we find reflexes with a raising contour tone before glottal stop, which could perhaps be seen as a sub-type of the high tone reflexes. In Yucunani and La Batea Mixtec (Group 71), there is additionally a split between monosyllabic and bisyllabic reflexes before glottal stop: with the former, there is a high tone reflex, but with the latter there is a low-high contour reflex. Tepango Mixtec and Zacatepec Mixtec, the two varieties which retain the final glottal stop, do not fall into either of those groups.

A similar but not identical pattern is found with reflexes of Proto-Mixtec *low, see Figure 4.5. Most varieties have a low tone reflex, but Jicaltepec, Peñoles, and Coatzospam have a high tone reflex. There is also a set of varieties, mostly concentrated in the east, which seem to show a dissimilation, such that the reflex is mid in the first mora instead of the expected low. The glottal stop has largely the same effects on Proto-Mixtec *low as it did on Proto-Mixtec *high, except that we cannot see the lowering effect, since all of the varieties that would exhibit it have a low tone reflex anyway. The varieties which exhibit an effect of the lost final glottal stop are all in Group 7 (Josserand's (1983) 'Baja' region), while those that do not fall into Groups 1-6 (Josserand's (1983) 'Alta' region). The raising vs. lowering effect does not exactly separate varieties along lower-level subgroups, but there are definitely tendencies worth exploring further. In Group 7.1 (Josserand's (1983) Mixtepec area) for example, all varieties show a raising effect, while all varieties in Group 7.2 and Groups 7.4-7.6 (Josserand's (1983) Central and Western Baja and Tezoatlan areas) show a lowering effect. Group 7.3 (Josserand's (1983) Guerrero and Southern Baja area) is split, with the majority of reflexes showing a raising effect, but a minority a lowering effect.

4.3 Methods

Phylogenetic signal and rates of change can only be calculated based on a tree or set of trees. I use the posterior distribution of trees from a previous Bayesian phylogenetic analysis of Mixtecan with BEAST2 (Bouckaert et al. 2019), which was based on annotated cognate sets from 130 varieties (for details see Auderset et al. submitted). I pruned the posterior distribution of trees to only contain the 42 languages for which I have adequate tonal data. From this, I took a sample of 1000 trees (sampled with LogCombiner from BEAST2) for further analysis. The maximum clade credibility (MCC) tree of the sample languages is presented in Figure 4.2. This is the most up to date family tree available, as previous classifications of

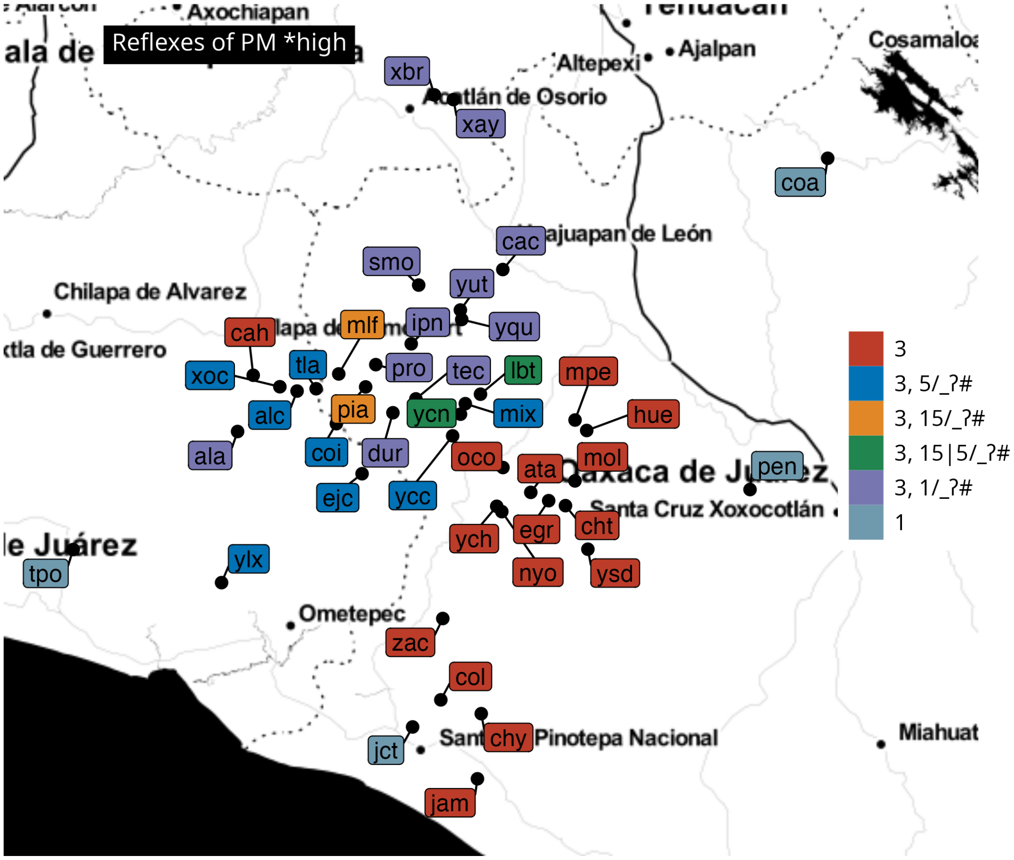


Figure 4.4: Reflexes of Proto-Mixtec *high in basic sets

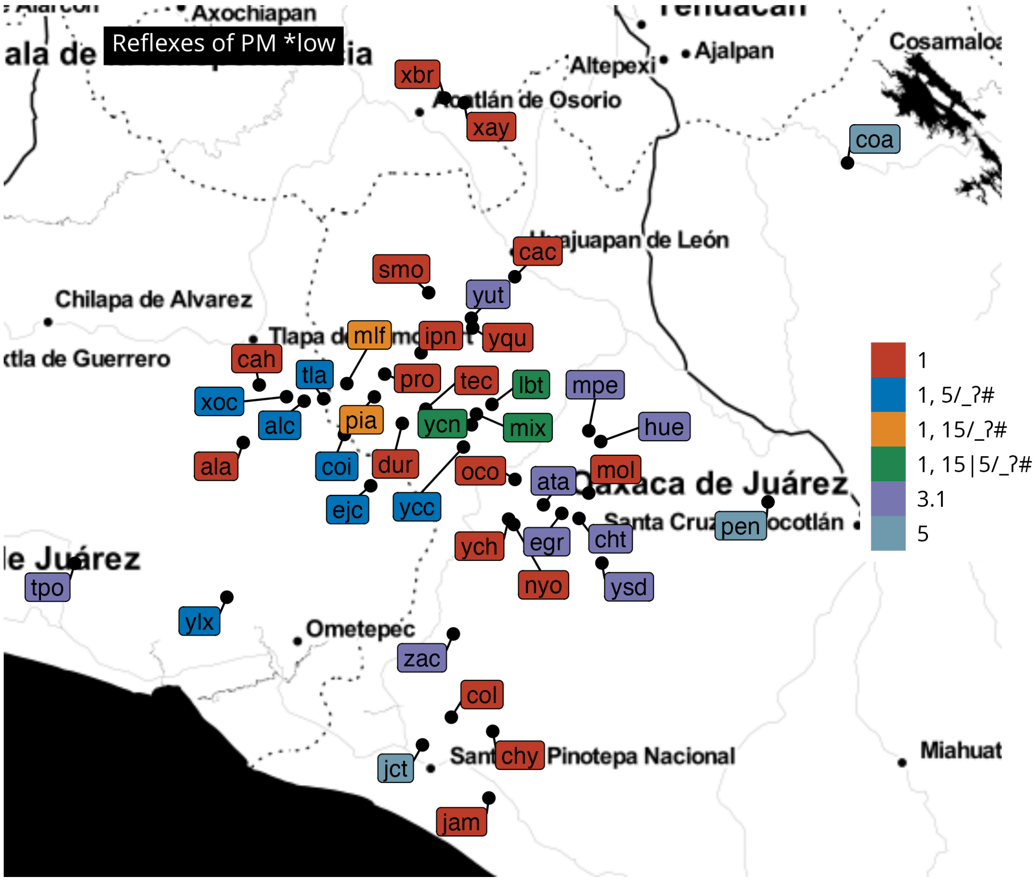


Figure 4.5: Reflexes of Proto-Mixtec *low in basic sets

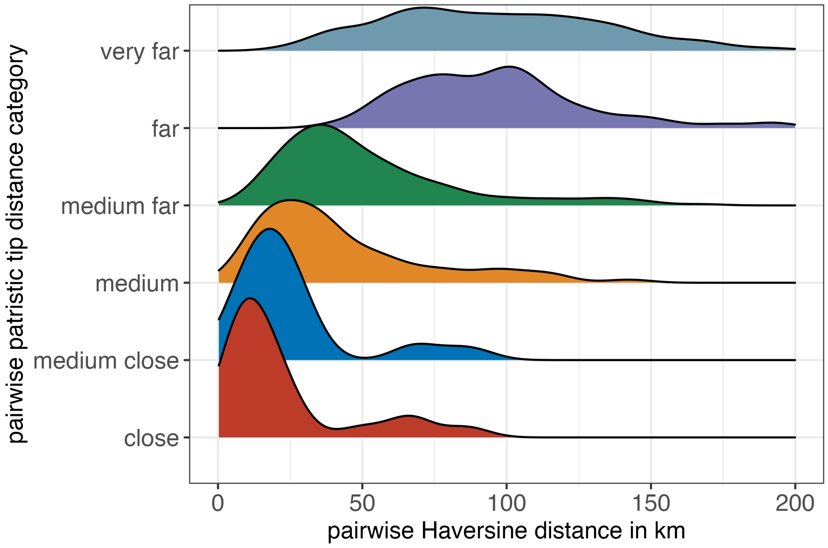


Figure 4.6: Pairwise geographic and phylogenetic distance plotted against each other (distances over 200km not shown because of very low density)

Mixtec relied on dialect areas based on isoglosses (Josserand 1983). Given that the Mixtec languages are described as a dialect continuum, one could argue that geographic distance should largely reflect phylogenetic distance as well. However, the Mixtec peoples are also known for a long history of migration, for example to the coast and the far northwest of Oaxaca state (Josserand 1983). Indeed, geographic distance does not correlate with phylogenetic distance as straightforwardly as one might expect. Figure 4.6 shows pairwise geographic distances plotted against phylogenetic distances. One can observe a general tendency for languages that are closely related to be geographically close and *vice versa*. However, languages that are geographically further away from each other (50km or more), can also be closely or relatively closely related to each other.

I measure phylogenetic signal with the metric D for binary traits (Fritz & Purvis 2010) calculated with the function *phylo.d* from the R package *caper* (Orme et al. 2018). This metric is based on the sum of sister-clade differences across a given tree. It provides a measure for the tendency of related species (sisters in the tree) to be more similar to each other than to other species sampled randomly from the tree (Fritz & Purvis 2010, Münkemüller et al. 2012). This reflects the tendency of closely related languages to be more similar to each other with respect to a given variable than to more distantly related languages in a given language family tree. We calculate the *D*-metric for each sound change for each tree from the posterior sample and aggregate the results per variable. The metric is interpreted with respect to two anchor points: if a trait, here a sound change, is the same across sister languages *D* will be 0, indicating phylogenetic signal; if a trait has completely different values in sister languages *D* will be 1, indicating low phylogenetic signal (Fritz & Purvis 2010:1043-1044). Values between 0 and 1 or outside of this range are interpreted with respect to these anchor points.

I measure evolutionary rate with a Hidden Markov Model as implemented in the function *corHMM* from the R package *corHMM* (Beaulieu et al. 2021). Since we do not expect the rates of gain and loss of sound changes to be the same, I used the ARD (= all rates differ) setting, which allows them to vary. The model estimates transition rates between the presence and absence of a sound change across a tree from the posterior sample. The 'hidden' part of the term refers to the fact that in this type of model multiple processes can be invoked to describe the evolution of the observed characters (Beaulieu et al. 2013). There are two states (presence and absence) and two rate categories (fast and slow) and the model assigns equal probability to a feature for belonging to one of these categories. This means that there are eight possible transition rates, from each state-rate combination to all others. Those transition rates cannot be observed directly but only derived from the states, therefore the term 'hidden' (Beaulieu et al. 2013:726). I follow Maddison et al. (2007) and FitzJohn et al. (2009) in the settings of the root prior. To assess the correlation – or absence thereof – of phylogenetic signal and evolutionary rate, I calculated the Kendall rank correlation coefficient (Kendall's τ) with the function *cor.test* in R (R Core Team 2022).

4.4 Results

4.4.1 Phylogenetic signal in tonal and segmental changes

To analyze and interpret the phylogenetic signal of tone changes on the one hand and segmental changes on the other, I summarized *D*-metric and standard deviation for each sound change variable across the 1000 trees by taking the median. For ease of interpretation, I classified the median *D*-metric values into four broader categories, based on the anchor points mentioned above: a median *D*-value higher than 1 is classified as overdispersed, values between 0.5 to 1 as random, values 0 and 0.5 as showing brownian motion (phylogenetic signal), and values under 0 as showing strong phylogenetic signal. Each sound change variable thus falls into one of these four categories based on their median *D*-metric. Some of the variables exhibit a very large standard deviation. I set the threshold for a 'high' standard deviation at slightly above the maximum absolute value of phylogenetic signal, i.e. at 9 (see the supplementary materials for more details). A closer look at the variables with such high standard deviation reveals that most of them are absent in all but one language. While a such a sound change is highly informative for the one language it appears in, it is not informative for the internal structure of the language family as a whole. I thus exclude all variables that have a very high standard deviation from further analysis. This affects 7 variables in total (4 pertain to consonants, 2 to vowels, and 1 to tones).[3]

Over three quarters (78%) of the 186 sound changes analyzed show phylogenetic signal and over half (61%) exhibit strong phylogenetic signal. This indicates that the sound changes identified (and the coding schema implemented) correlate well with the cognacy data underlying the phylogenetic trees. This might seem obvious, but given the characterization of the Mixtec language family as a dialect continuum which violates the standard assumptions of the comparative method and should thus not be represented as a tree (Bradley & Josserand

[3] Consonant variables: K02, W09, W12, X04; Vowel variables: E18, U10; Tone variable: Tone56.

1982:303) it is an important finding. It adds further support to the view advocated here that even if parts of the family are not tree-like, the comparative method applies just as in other families. Table 4.9 provides an overview of *D*-metric categories by type of sound change variable. Vowels have the most variables with strong phylogenetic signal at almost three quarters. Consonants and tones have less than that, but still over half of the variables have strong phylogenetic signal. Tones do exhibit a slightly lower percentage than the other two categories, but the majority of tone change variables (75%) still show phylogenetic signal.

Table 4.9: Percentage of variables per *D*-metric category per sound change type with absolute numbers given in brackets (variables with very high standard deviation excluded)

D Category	Consonants		Vowels		Tones	
strong phy. sig.	57%	(56)	72%	(57)	52%	(29)
brownian	17%	(17)	11%	(9)	23%	(13)
random	12%	(12)	5%	(4)	14%	(8)
overdispersed	14%	(14)	11%	(9)	11%	(6)

The distribution of median *D*-values across the three categories of sound change is illustrated in Figure 4.7. The mean and standard deviations of the three categories are summarized in Table 4.10 and for each variable separately in the supplementary materials. Tones have a higher mean and median than both consonants and vowels, but the lowest standard deviation, see Table 4.10. Taking into account the standard deviations, the three sets of sound changes are not significantly different from each other, or in other words, they all overlap. This is reflected in the density distributions as well, which also show that all three categories have a small number of variables with very high phylogenetic signal (<-8). Vowels also have a small number of highly overdispersed variables, which we do not find in consonants, and only to a much lesser extent in tones. All of these overviews and summary statistics reflect relatively small differences between tones and segments with respect to phylogenetic signal. It is certainly not the case that tone change variables overall are distributed randomly across varieties and carry no phylogenetic signal. If anything, vowels seem to behave differently

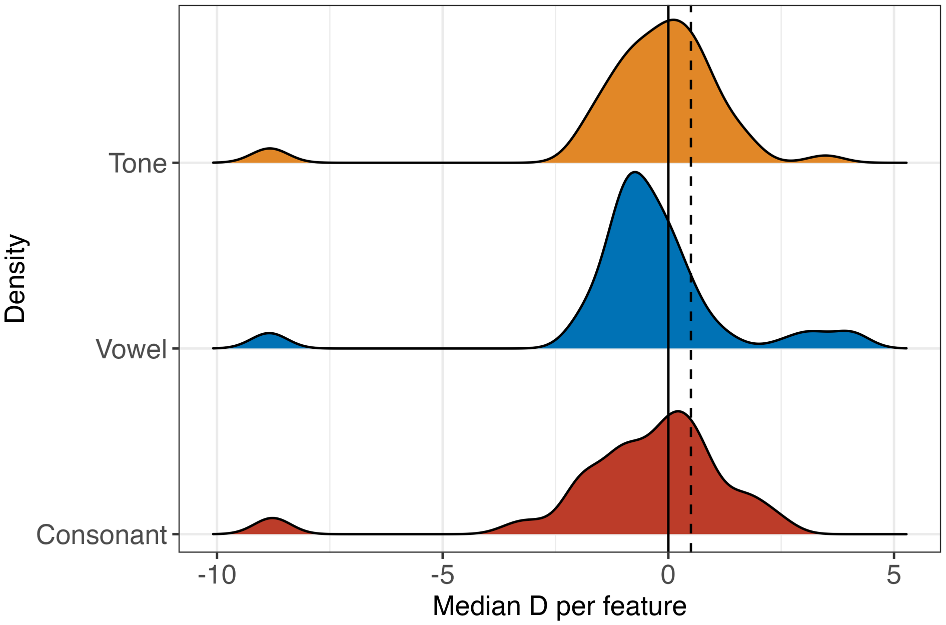


Figure 4.7: Density of median *D* per sound change type with line at 0 and dashed line at 0.5

from consonants and tones, showing a stronger phylogenetic signal overall.

Table 4.10: Mean, median, and standard deviation of *D* per sound change category (variables with very high standard deviation excluded)

Type	Mean	Median	SD
Consonants	-0.59	-0.14	2.12
Vowels	-0.50	-0.62	2.14
Tones	-0.37	-0.12	1.94

Of course, the variables within each of these aggregated categories do not behave uniformly. I thus explore consonantal, vowel, and tonal variables in more detail below. Figure 4.8 shows the distribution of *D* in consonant variables split according to the manner of articulation.[4] This is of interest because the glides **j* and **w* undergo many changes and are often lost (Bradley & Josserand 1982) and could thus be expected to show less phylogenetic signal than plosives or fricatives. They also constitute the conditioning environment for many

[4] Nasal and glottal stop variables are excluded because they are so few in number that they cannot be displayed on this plot.

other sound changes (Auderset & Campbell in prep). Glides indeed have the highest density slightly above zero (median D=0.01, SD=0.74), but this is also the case for fricatives (median D=-0.07, SD=0.92), while plosives (median D=-0.95, SD=1.14) have their peak below zero. While all three categories exhibit some variables with very strong phylogenetic signal, only glides have very overdispersed variables. To sum up, variables related to the glides do not as a whole behave much differently from other consonant variables, but they do show more variables within their category that have no phylogenetic signal. There are four consonant changes that exhibit very strong phylogenetic signal (D<-8) and a further three with strong phylogenetic signal (D<-3):

GS03	*ø > ʔ /_C[sonorant]	in Group 1 (Coatzospam)
T07	*t > ts /_ɨ, ij	in Group 1 (Coatzospam)
J14	*j > ø /_e	in Group 1 (Coatzospam)
ND09	*^{n}d > ^{n}z /_ɨ	in Group 74
T12	*t > s /_ɨ	in Group 74
T21	*t > t^{n}/_ĩ	in Group 3

All of the consonantal changes with strong phylogenetic signal separate out specific groups and most of them concern changes of plosives before front vowels and *ɨ*. The first three are specific to the Group 1 variety of Coatzospam, setting it apart from all other varieties. A further two changes delineate Group 74 from other varieties in Group 7 and beyond, while one change appears only in Group 3. I did not systematically investigate all consonant changes with phylogenetic signal in this way, but the changes illustrated above already provide good diagnostics for assigning varieties to these respective groups.

Mixtec vowels are often divided into two groups: the weak or 'inner triangle' vowels *e, ɨ, o* and the strong or 'outer triangle' vowels *a, i, u.* The triangle terminology is based on the common representation of the vowel space as a triangle or trapezoid. The strong vs. weak labels

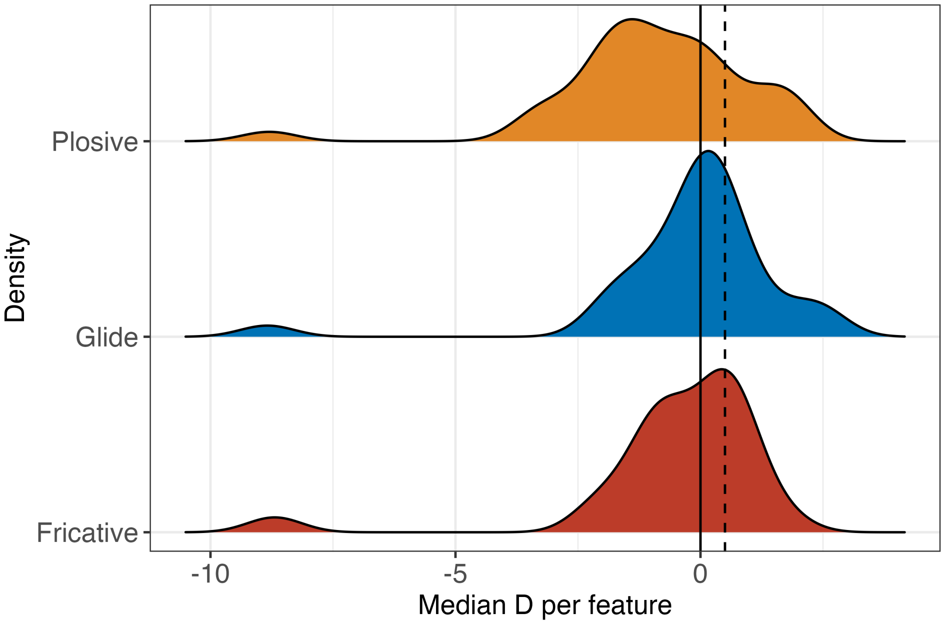


Figure 4.8: Density of median *D* per type of consonant with line at 0 and dashed line at 0.5

are based on the impression that weak vowels undergo more changes, i.e. are less stable, than strong vowels and on the observation that they are less frequent in contemporary varieties (Josserand 1983:245-250). Figure 4.9 shows the distribution of median *D* in vowels split into these two categories. While both inner and outer vowel changes have their peak below zero, outer vowels have a small number of variables that are very overdispersed, which we do not find in inner vowel variables. Conversely, all variables with very high phylogenetic signal are found in inner triangle vowel variables. To sum up, there are more outer vowel variables (median D=0.06, SD=1.16) that show no phylogenetic signal than inner vowel variables (median D=-1.43, SD=0.79). This runs counter to the statement mentioned above that outer triangle vowels are more stable, or at least it raises questions for further research. We could expect that stable vowels exhibit stronger phylogenetic signal than the 'inner triangle' vowels, but it seems that the changes in these vowels mostly fall along the lines of subgroups. There are three vowel variables that show very strong phylogenetic signal with a *D*-value under -8 and one with a high phylogenetic signal under -2. They all pertain to the Proto-Mixtec vowel **e*,

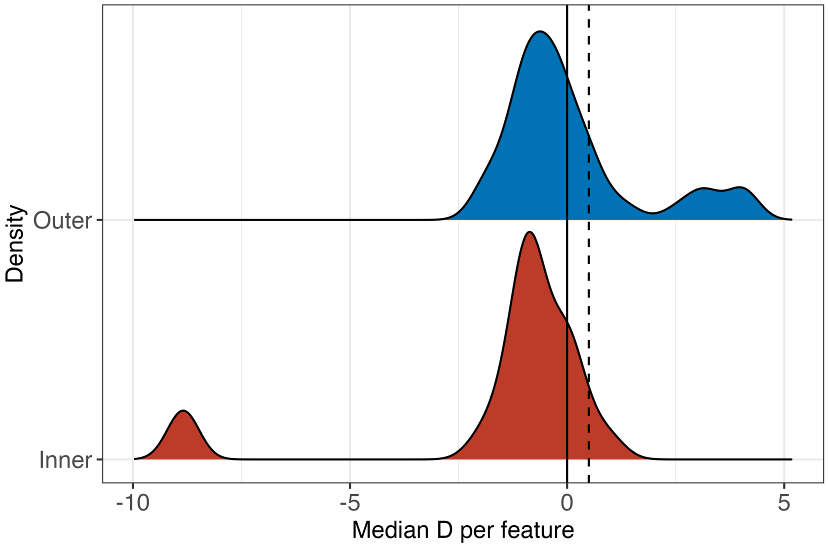


Figure 4.9: Density of median *D* per type of vowel with line at 0 and dashed line at 0.5

which is known to undergo many changes and in some varieties changes to *a* in almost all contexts (Auderset & Campbell in prep):

E04	*e>a /_xĩ	absent in Group 1 (Coatzospam) and Yucuhiti and Nuyoo (Group 6)
E17	*e>a /ux_	absent in Group 1 (Coatzospam)
E03	*e>a /_ni, tĩ	absent in Group 1 (Coatzospam)
E01	*e>i /_(ʔ)ji	present in Group 1 (Coatzospam)

As opposed to the consonant changes with strong phylogenetic signal, these vowel changes separate subgroups and varieties mostly by being absent there, but present in all other varieties. Three of the changes delineate Group 1 variety Coatzospam from all others (either by presence or absence), while a forth change is also absent from the closely related Group 6 varieties of Nuyoo and Yucuhiti.

As explained in section 4.2.2, tone melodies in Mixtec are usually classified into two groups, 'basic' and 'modified' (Longacre 1957). To recap, modified means that the tone melody contains a tone restricted to tone sandhi environments and is said to have an unpredictable

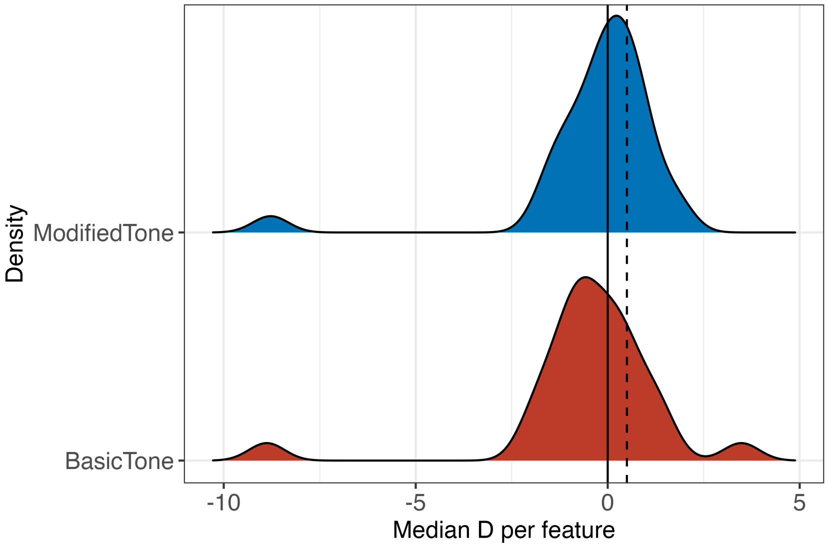


Figure 4.10: Density of median *D* per type of tone set with line at 0 and dashed line at 0.5.

reflex, while basic tone melodies contain only reflexes of the Proto-Mixtec *high and *low tone with or without final glottal stop. Figure 4.10 shows the distribution of *D*-values in basic and modified tone sets. Basic tones (median D=-0.49, SD=0.73) have the peak below zero, while modified tones (median D=-0.27, SD=0.53) have the peak just under 0.5. Both categories have a few variables with very strong phylogenetic signal, but only basic tones have a small number of highly overdispersed variables, as can be seen in Figure 4.10. While there are small differences between these two sets of tones, we can definitely say that modified tones show no more or less phylogenetic signal than basic tones. This suggests that in terms of diachronic change, 'modified' tones operate much the same as basic tones.

I also investigated potential differences between reconstructed Proto-Mixtec tones *high and *low. The phylogenetic signal is a bit stronger with Proto-Mixtec *low (median D=-0.56, SD=0.70) than *high (median D=-0.19, SD=0.56), but the difference is minor. There are two tone variables that show very strong phylogenetic signal (D <-8), while other tonal variables are all under -2, two of them relatively close to that (D ~-1.9), see Figure 4.11:

Tone22	*L >15 /L_#	in Group 1 (Coatzospam)
Tone34	*L >15 / (B5)_#	in Piedra Azul, SM la Flor, Tlahuapa, Yoloxochitl (Group 73)
Tone09	*H >15 /_ʔ#	in Piedra Azul and San Marcos la Flor (Group 73)
Tone35	*H >1 /_L	in Jamiltepec (Group 2) and Tepango (Group 73)

Tone22 is only found in Group 1, along with many other variables with strong phylogenetic signal already discussed. Tone09 neatly separates out the closely related varieties of Piedra Azul and San Marcos la Flor, where the lost final glottal stop turns Proto-Mixtec *high into a low-high contour tone. Tone34 also includes these varieties, but also Tlahuapa, which is quite closely related to them as well and Yoloxochitl, which is a bit more distant. Tone35 is limited to two varieties of different groups.

Figure 4.11 reveals interesting differences with respect to the tone sets. In Set A, where the basic set is reconstructed as *HH(ʔ), most tone change variables show phylogenetic signal – some even quite strong phylogenetic signal – and this is true for changes in modified as well as basic sets. A similar distribution is found in set D (basic set reconstructed as *LH), which has no random or overdispersed variables at all. Conversely, sets B (basic set reconstructed as *LL(ʔ)) and C (basic set reconstructed as *HL) both have variables with a wide range of *D*-values. With the data available as of now and the lack of detailed phonetic studies, I cannot explain the discrepancy between the tonal changes in these sets. However, these questions open avenues for further research, which should investigate the diachrony of each of those changes in more detail.

4.4.2 Rates of gain and loss in tonal and segmental changes

I calculated the rate of gain and loss for each sound change variable across the 1000 trees. The resulting rates cannot be interpreted in absolute terms, but only relative to each other because the phylogenetic trees cannot be anchored in real time due to absence of good cali-

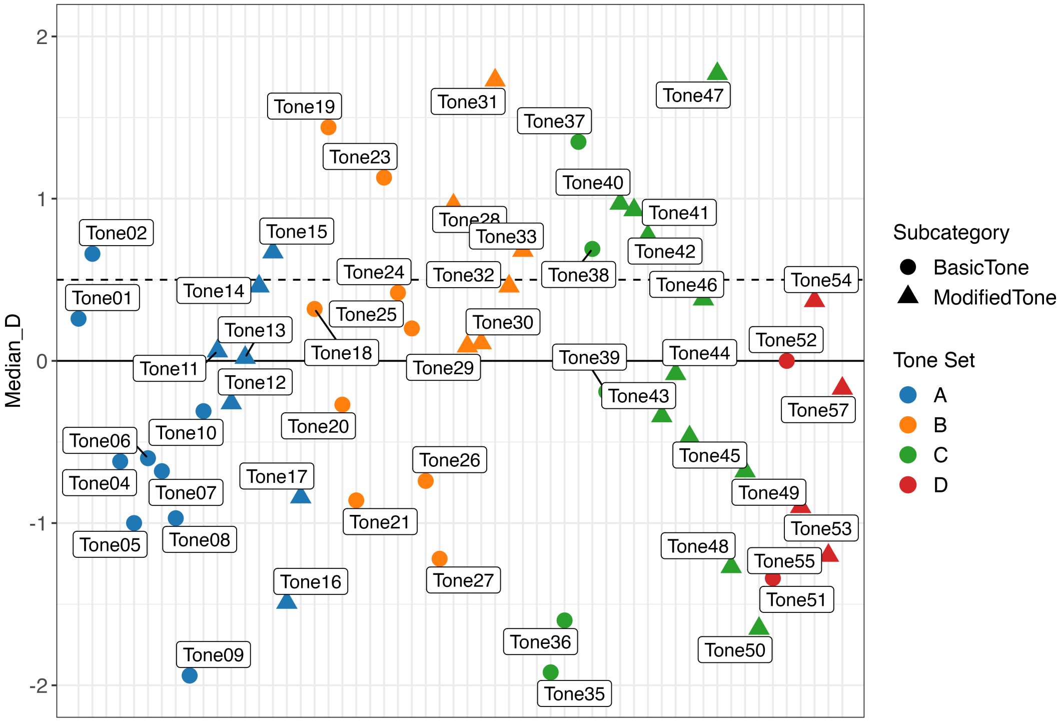


Figure 4.11: Median *D* of tone variables colored by tone set with line at 0 and dashed line at 0.5 (Variables Tone22 and Tone34 at around -8 and Tone03 at 3.5 excluded for better visibility of other variables. Full plot in the supplementary materials.)

bration points. I log-transformed the rates for better visualization and easier interpretation. I excluded variables that show a high standard deviation in both the rate of loss and gain. I set the threshold for this at 0 (for the log-transformed SD), since no variables show a non-negative log-transformed rate. This excludes 4 variables in total, 2 consonant variables, and 2 tone variables.[5] Figure 4.12 summarizes the results split into consonant, vowel, and tone variables on rates of gain and loss. (Histograms of the raw, unaggregated values can be found in the supplementary materials.) Values that are close to zero indicate a fast rate of change (raw value just below 1) and values close to -10 a very slow rate of change (raw value approaching 0). In all three categories – consonants, vowels, and tones – the rate of gain is overall slower than the rate of loss, but the difference is not large. Furthermore, the median rates of gain and loss across the three categories are similar and tones do not stand out against consonants and vowels in this respect either. In all three groups, there are some variables with rates around -10, which means that they do not change at all, or that they change extremely slowly. The majority of variables in all three groups appear around the median rate value. This means that variables are overall gained and lost, respectively, at relatively similar rates. Vowels have a few variables which are lost quite fast (rate >-1), but in tones and consonants we do not find such a fast rate of loss to the same extent. Consonants also have a small number of variables that are gained fast, a pattern absent in vowels and tones.

There is a tendency to equate low phylogenetic signal with a high rate of change and conversely high phylogenetic signal with a low rate of change (Revell et al. 2008). Framed in linguistic terms, structural variables that are very stable are often assumed to be good candidates for establishing relationships between languages or language families (Nichols 2003, Dediu & Levinson 2012). However, the relationship between these two processes is more complex than that and they are not usually correlated in biology (Revell et al. 2008).

To assess whether there is a correlation between phylogenetic signal and rates of gain

[5] Consonants: J03, J08; Tone: Tone28, Tone29.

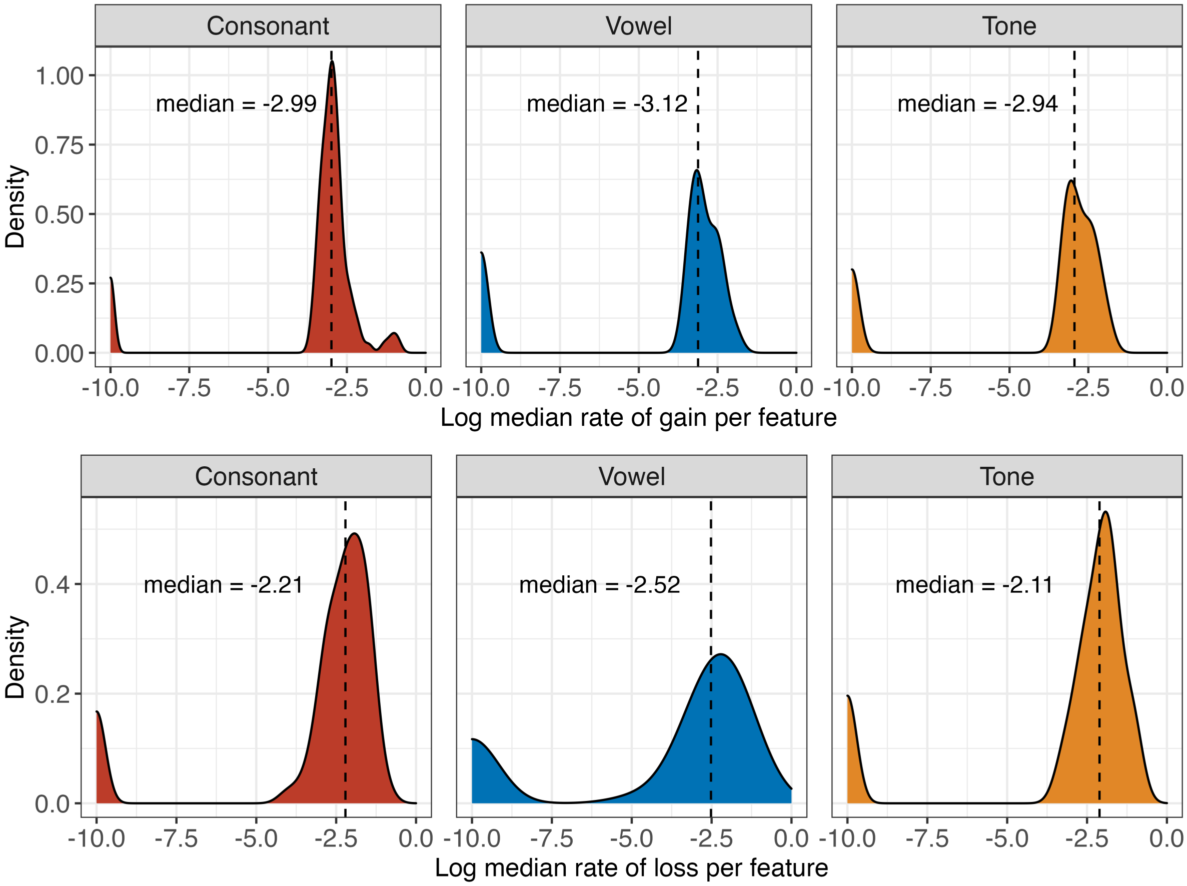


Figure 4.12: Median rates of gain and loss per variable aggregated per variable with ARD. Dashed lines indicates group medians.

and loss in sound changes, I computed the Kendall rank correlation coefficient for these two measures. The results are summarized in Table 4.11. There is a moderately strong positive correlation between phylogenetic signal and both the rate of gain (τ=0.38) and loss (τ=0.44) of sound changes. However, consonants, vowels, and tones behave quite differently with respect to this correlation. In consonants, there is little difference between loss and gain, the latter being just a bit lower. Vowels show a quite strong correlation of the rate of loss with phylogenetic signal, but conversely a weaker one with the rate of gain. Tones exhibit the exact opposite pattern, although the difference is not as stark: here the rate of gain is more strongly correlated with phylogenetic signal than the rate of loss.

The overall correlation is not as strong as the one found in Hübler 2022 with structural variables in five Eurasian language families (τ=0.51 for loss and τ=0.50 for gain), but it is interesting that it occurs in a completely different language family based on sound change variables. As mentioned above, in biology this correlation between phylogenetic signal and rates of change is not generally found. It is possible and should be investigated further whether the correlation between phylogenetic signal and rates of change is perhaps a characteristic of linguistic evolution that sets it apart from biological evolution. In fact, if it could be shown that a slow rate of change and strong phylogenetic signal correlate cross-linguistically and across different components of language, this would make a strong case for further investigating deep relationships beyond the comparative method.

Table 4.11: Correlations between phylogenetic signal and rates of gain and loss (Kendall's tau)

Category	Loss		Gain	
	τ	p	τ	p
overall	0.44	<0.01	0.38	<0.01
consonants	0.48	<0.01	0.46	<0.01
vowels	0.50	<0.01	0.27	<0.01
tones	0.32	<0.01	0.44	<0.01

Intuitively, we also expect that variables which are present in almost all languages have a high rate of gain and a low rate of loss, and conversely, variables that are absent in almost all languages have a low rate of gain and a high rate of loss. Figure 4.13 plots the rate of loss and gain against phylogenetic signal indicating in how many languages each variable is present across the whole data set. We do see that the variables with the highest rates of gain are present in the majority of languages, while those lost at the highest rates are absent in most languages. The variables present in almost all languages do indeed generally exhibit a relatively low rate of loss and a high rate of gain. Since many variables are absent in the majority of languages of the sample, this distribution is less clear. However, the plot also reveals that the relationship between presence/absence of a variable and its rate of change is not a simple linear correlation. To explore this relationship further, I calculated Kendall's tau for the presence of each variable with phylogenetic signal, rate of loss, and rate of gain. There is a moderately strong correlation with the rate of gain (τ=-0.32, p=<0.01), a moderate correlation with the rate of loss (τ=-0.21, p=<0.01), but only a very weak correlation with phylogenetic signal (τ=0.14, p=<0.01). We can explain the latter in terms of mechanisms of sound change more broadly. Both sound changes that are very frequent and very rare within a family could be innovations of closely related varieties or subgroups and thus help resolve the family tree. This is what is reflected in the weak correlation of presence of a variable with phylogenetic signal.

Since this study is primarily concerned with comparing rates of change in tones versus segments, Figure 4.14 shows the same rates against phylogenetic signal plot, but this time colored by sound change type. Based on the recurring statements in the literature that tones change faster than segments (see Section 4.1 for references), I expected to find faster rates of gain and loss for tones than for vowels and consonants. This is not borne out by the data, as variables of all three types are spread out over slow and fast rates of gain and loss.

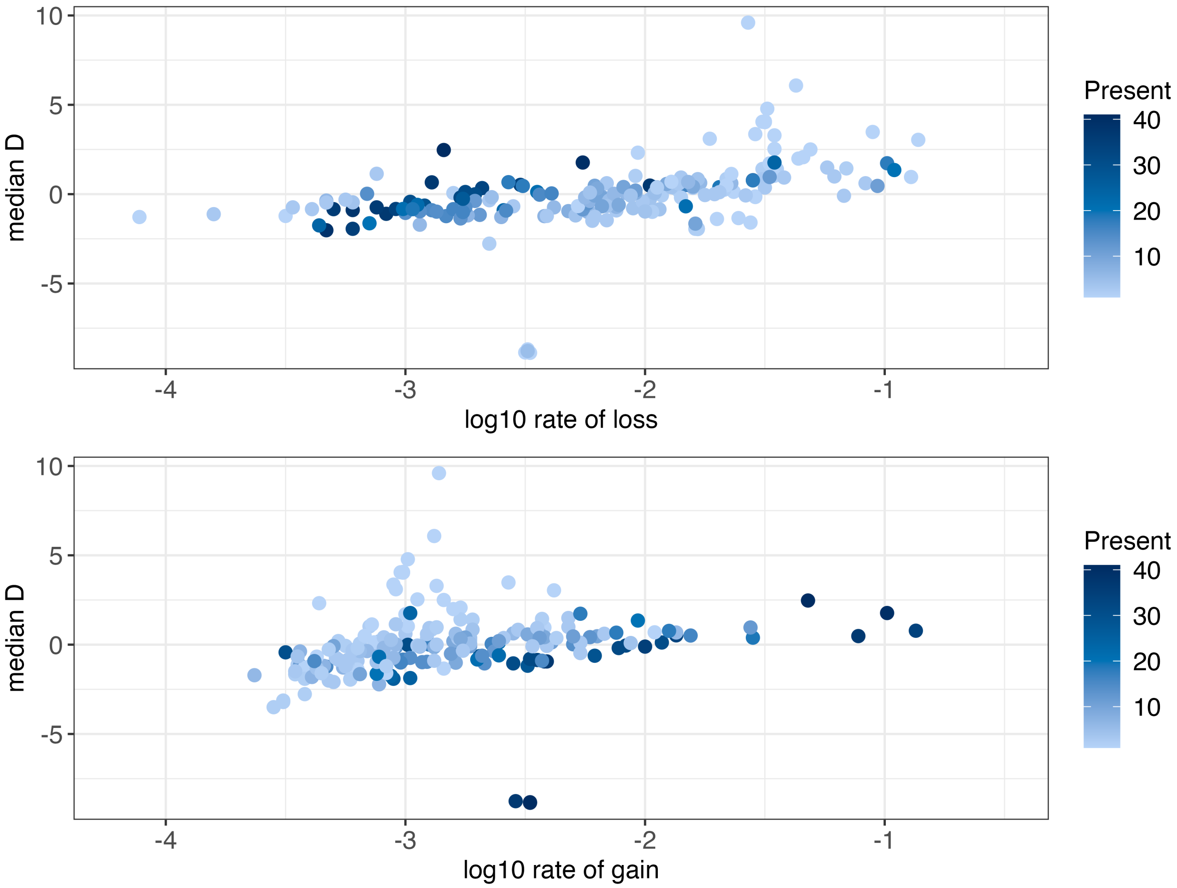


Figure 4.13: Rate of loss and gain (log10) against *D* colored by number of languages in which a variable is present

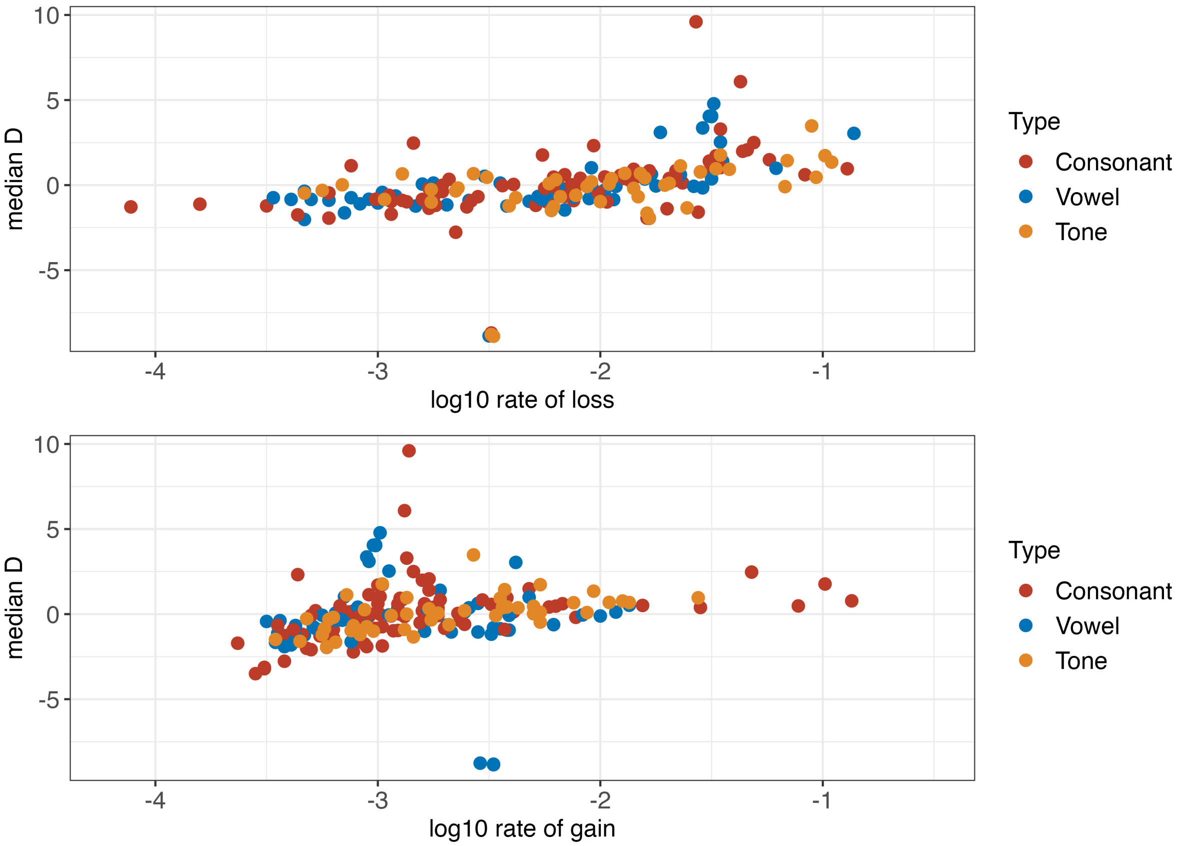


Figure 4.14: Rate of loss and gain (log10) against *D* colored by type of variable (languages that show no change are excluded from the plot)

4.5 Discussion

In the previous sections, I explored phylogenetic signal and rates of change in tone across 42 Mixtec languages. Based on the metrics calculated, tones do not generally change faster than segments, nor do they show less phylogenetic signal in this language family. I found a moderately strong positive correlation between phylogenetic signal and rates of loss and gain. This indicates that sound changes which proceed more slowly in general have stronger phylogenetic signal than those that proceed at a faster rate. This is true of both tones and segments. The relationship between the presence of a sound change across the sample languages with phylogenetic signal is, as expected, more complex, such that both changes which are frequent and those which are rare can contribute to subgrouping.

There are certain limitations to the study that invite follow-up studies of similar scope in the future. First, due to lack of good descriptions of tone systems, our sample of Mixtec varieties is somewhat limited and does not cover all previously identified subgroups to the same extent and excludes one subgroup (Linkage 5) completely. Of two groups (Group 1 and 4), I was only able to sample one variety each and thus we cannot know how representative these varieties are of their respective subgroups. The same applies within the large Group 7, where the sample does not cover all subgroups to the same extent. Second, the reconstruction of tone and the identification of tone changes similarly suffers from this lack of coverage. While I believe that the reconstruction proposed by Dürr (1987), confirmed by Swanton & Mendoza Ruíz (2021) and the data used in this study is in broad strokes correct, the 'modified' sets are still in need of revision, especially with respect to the potential environments that conditioned them. Third, the sound changes were identified and coded based on the same word list that served as the basis for the cognate sets used to construct the phylogenetic trees. The reason for this is a practical one, since no other comparative Mixtec data on such a large scale is available and assembling such a data set is an undertaking of considerable time and effort outside the scope of the present study. However, I also believe that this does not invalidate the results. Sound changes are regular and should operate the same across the lexicon, since we assume that the relationship between form and meaning is (largely) arbitrary (Rankin 2003:184). There are parts of the lexicon in which this assumption is violated, for example where we find sound symbolism or other forms of iconicity. However, such forms are excluded when creating cognate sets, precisely because they often display various irregularities. This means that the sound changes identified based on the cognates sets used for the phylogeny are in principle identical to those we would find based on a different non-overlapping set of cognates. In practice, they of course only constitute a subset of all sound changes that took place in Mixtec and future studies might identify additional changes or lead to modifications of existing ones. But this is not a consequence of the non-independence from the earlier data set, but rather of the practical necessity to work with a subset of the lexicon.

4.6 Conclusion

Historical comparative linguistics is an important tool for furthering our understanding of the past not least because it provides a resolution and granularity that is often not possible to achieve in archaeology and genetics (Kaufman 1990:14-15). Tone is an integral feature of Mixtec languages, as well as many other languages of Mesoamerica and beyond. Despite its prevalence, tone has received limited attention in historical linguistics, both with respect to traditional and quantitative methods. As a consequence, our knowledge of tone change processes as opposed to those of segmental change is still limited. This study presents a first step towards filling this *lacuna* by making some of the assumptions with respect to tone change explicit and testing them with quantitative methods based on a carefully curated data set. As I have shown, tone change in Mixtec does not behave differently from segmental change in any significant way. Many tone changes carry phylogenetic signal and can thus contribute to our understanding of the internal structure of this language family just like segmental changes. Tones also do not change faster or slower than segments overall, exhibiting similar transition rates as segments. These two measures suggest that tone change operates much the same way as segmental change and should be investigated on a par with segmental change. Even though this study is limited to one language family, the methodology presented here can easily be expanded to other language families, and it is indeed my hope that it will inspire follow-up studies to close the tone gap in historical linguistics in the near future.

4.7 Data availability

The data, scripts, and other supplementary materials can be found at `https://osf.io/ts4kw/?view_only=a01508a7d2474c19a20576bdc9df6980`.

Chapter 5

Conclusion

The overarching goal of the three chapters presented in this dissertation was to bring Mixtec historical linguistics to the twenty-first century by revisiting earlier research in light of new data and methods and by operationalizing and addressing novel questions. The results have implications for our understanding of the linguistic history of Mixtec languages, but also contribute to the field of historical linguistics more broadly. Underlying all three studies is a carefully compiled and annotated database of over 1000 cognate sets covering 137 Mixtecan languages. With its standardization of orthography to IPA including tone notation, it will be a valuable tool beyond this dissertation and can serve as a model for other language families in Mesoamerica and beyond.

In chapter 1, I showed that subgrouping is possible and informative, despite the characterization of Mixtec as a dialect continuum. Through the application of Bayesian phylogenetics, I was able to identify more tree-like and more wave-like parts of the language family and quantify the levels of uncertainty with respect to each grouping. The results are congruent with what we already know about the migration history of Mixtec peoples, reflected for example in the clear split of the coastal varieties from the rest, but the results also open avenues for further research. In chapter 2, I showed that the patterns and distributions of segmental sound changes in Mixtec largely align with the subgroups proposed through the Bayesian phylogenetic study based on cognacy data and to a lesser extent also with dialect areas previously established through overlapping vowel isoglosses. These studies taken together indicate that the dynamics of sound change and lexical replacement in 'dialect continua' are not fundamentally different from those in other language families. We can apply the same methods and tools to investigate their linguistic history and our focus should lie on integrating the results with other disciplines such as archaeology and anthropology to gain a deeper understanding of the processes at work and the finer details of the prehistory of the people, their language, and culture. Likely the situation found in the Mixteca is not inherently different from that of other, better studied language families.

In chapter 3, I addressed the question of whether tones change faster than segments in Mixtec and showed that there are no substantial differences in rates of change or phylogenetic signal between segments and tones. This empirical study thus contradicts anecdotal assumptions on the volatility and instability of tones and calls for an end to the resistance to including tone in historical linguistics. It also provides a framework for diachronic comparative tone studies beyond tonogenesis in other language families. The integration of computational and quantitative methods in historical linguistics has significantly advanced the field, but has also been met with resistance based on a perceived opposition to qualitative, traditional approaches. However, rather than competing with each other, these two broad approaches work in tandem and the best results can be achieved by a rigorous application of both, as in this dissertation.

In each of the three studies, I also identified avenues for further research. One of those concerns the crucial role of good language documentation in historical research on language families where we have little or no access to earlier forms of the languages. One desideratum for future research is thus another large-scale language survey carried out in the Mixteca. The last such survey was conducted in the Mixteca region in the late 1970s, focusing on

Mixtec, but an updated survey covering Triqui and Cuicatec varieties as well would put us in a better position to answer outstanding questions on the relationship of the higher-level branches of Mixtecan. Another avenue for further work would be to apply these same methods to, for example, the Zapotecan language family, a similarly highly-diversified branch of Otomanguean, that also continues to receive considerable attention in language documentation and description. With respect to the sound change data, detailed smaller scale studies investigating convergences and divergences with the lexical data would lead to testable hypotheses of migrations that could then be evaluated in light of archaeological work. The study of tone could be advanced significantly, if we had more descriptive materials that include tone systems. To achieve this, we need to shift away from viewing tone as negligible or secondary, especially in languages like Mixtec where it carries a high functional load in lexicon and grammar. Finally, quantitative methods such as phylogenetics have to date only been applied to a relatively small number of language families, often focusing on large, well-studied families. I believe that applying these methods to a more diverse set of language families, which also entails a more diverse set of linguistic histories, will improve not just the methodology but also our general understanding of linguistic evolution in the future. In addition, these advances can also help improve the creation and dissemination of community materials. If we better understand the internal relationships of these languages, we can make more informed decisions about sharing materials between communities that really understand them and support documentation efforts in a more targeted way.

Bibliography

Alexander, Ruth Mary. 1980. *Gramática mixteca de Atatlahuca.* México, D.F.: 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.

Anonymous. 2022. Supplementary Materials to "Subgrouping in a 'dialect continuum': A Bayesian phylogenetic analysis of the Mixtecan language family". https://doi.org/10.5281/zenodo.6513506.

Anttila, Raimo. 1989. *Historical and comparative linguistics.* Amsterdam: John Benjamins.

Arana Osnaya, Evangelina. 1960. Relaciones internas del mixteco trique. In *Anales del Museo Nacional de México* 12, 219–273. México, D.F.: Instituto Nacional de Antropología e Historia.

Auderset, Sandra & Eric W. Campbell. in prep. Patterns and distributions of sound change in Mixtec. *tbd* .

Auderset, Sandra, Simon J. Greenhill, Christian T. DiCanio & Eric W. Campbell. submitted. Subgrouping in a 'dialect continuum': A Bayesian phylogenetic analysis of the Mixtecan language family. *Journal of Language Evolution* tbd(tbd).

Beam de Azcona, Rosemary G. 2007. Problems in Zapotec tone reconstruction. In *Annual Meeting of the Berkeley Linguistics Society*, vol. 33 2, 3–15.

Bakker, Dik. 2010. Language Sampling. In Jae Jung Song (ed.), *The Oxford Handbook of Linguistic Typology*, chap. 6, 100–127. Oxford University Press.

Ballard, William Lewis. 1969. *Phonological history of Wu.* Berkeley: University of California Berkeley PhD dissertation.

Beaulieu, Jeremy, Brian O'Meara, Jeffrey Oliver & James Boyko. 2021. *corHMM: Hidden Markov Models of Character Evolution* r package version 2.7 edn. `https://CRAN.R-project.org/package=corHMM`.

Beaulieu, Jeremy M, Brian C O'Meara & Michael J Donoghue. 2013. Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms. *Systematic biology* 62(5). 725–737.

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

Beekes, Robert Stephen Paul. 1989. The nature of the Proto-Indo-European laryngeals. In *The new sound of Indo-European: essays in phonological reconstruction*, 23–33. Berlin: Mouton de Gruyter.

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

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

Bickel, Balthasar. 2007. Typology in the 21st century: major current developments. *Linguistic Typology* 11. 239 – 251.

Bickel, Balthasar. 2010. Capturing particulars and universals in clause linkage: a multivariate analysis. In Isabelle Bril (ed.), *Clause-hierarchy and clause-linking: the syntax and pragmatics interface*, 51–101. Amsterdam: Benjamins.

Bickel, Balthasar. 2015. Distributional typology: Statistical inquiries into the dynamics of linguistic diversity. In Bernd Heine & Heiko Narrog (eds.), *Oxford Handbook of Linguistic Analysis*, chap. 37, 901–924. Oxford University Press 2nd edn.

Bickel, Balthasar, Peter K Austin, Oliver Bond, David Nathan & Lutz Marten. 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*, vol. 3, 3–13. London: SOAS.

Bickel, Balthasar & Johanna Nichols. 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, MPI for Psycholinguistics Nijmegen.

Birchall, Joshua, Michael Dunn & Simon J Greenhill. 2016. A combined comparative and phylogenetic analysis of the Chapacuran language family. *International Journal of American Linguistics* 82(3). 255–284.

Bloomfield, Leonard. 1933. *Language.* New York: Holt.

Bouckaert, Remco, Timothy G Vaughan, Joelle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kühnert, Nicola De Maio et al. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS computational biology* 15(4).

Bouckaert, Remco R, Claire Bowern & Quentin D Atkinson. 2018. The origin and expansion of Pama–Nyungan languages across Australia. *Nature ecology & evolution* 2(4). 741–749.

Bouckaert, Remco R & Joseph Heled. 2014. DensiTree 2: Seeing Trees Through the Forest.

Bowern, Claire. 2013. Relatedness as a factor in language contact. *Journal of Language Contact* 6. 411 – 432.

Bowern, Claire. 2018. Computational phylogenetics. *Annual Review of Linguistics* 4. 281–296.

Bradley, C. Henry. 1968. A method for determining dialectal boundaries and relationships. *América Indígena* 28(3). 751–760.

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.

Bryant, David & Vincent Moulton. 2004. Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. *Molecular Biology and Evolution* 21(2). 255–265.

Byers, Douglas S (ed.). 1967. *The Prehistory of the Tehuacan Valley*, vol. 1: Environment and Subsistence. Austin, TX: University of Texas Press.

Cahill, Michael. 2011. Tonal diversity in languages of Papua New Guinea. SIL Electronic Working Papers.

Campbell, Eric W. 2017a. Otomanguean historical linguistics: Exploring the subgroups. *Language and Linguistics Compass* 11(7).

Campbell, Eric W. 2017b. Otomanguean historical linguistics: Past, present, and prospects for the future. *Language and Linguistics Compass* 11(4).

Campbell, Eric W. 2021. Why Is Tone Change Still Poorly Understood, and How Might Documentation of Less-studied Tone Languages Help? In Patience Epps, Danny Law & Na'ama Pat-El (eds.), *Historical Linguistics and Endangered Languages*, 15–40. Routledge.

Campbell, Eric W. & Griselda Reyes Basurto. forthcoming. El Tu'un Savi (mixteco) en California: documentación y activismo lingüístico. In Marcela San Giacomo, Fidel Hernández Mendoza & Michael Swanton (eds.), *Estudios sobre lenguas mixtecanas*, Seminario Permanente de Lenguas Mixtecanas, Instituto de Investigaciones Antropológicas, Universidad Nacional Autónoma de Méxic.

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

Carroll, Lucien Serapio. 2015. *Ixpantepec Nieves Mixtec word prosody*: University of California San Diego PhD dissertation.

Castillo García, Rey. 2007. *Descripción fonológica segmental y tonal del mixteco de Yoloxóchitl, Guerrero*. México, D.F. CIESAS MA thesis.

Chance, John K. 1986. 11. Colonial Ethnohistory of Oaxaca. In *Supplement to the Handbook of Middle American Indians*, vol. 4, 165–190. University of Texas Press.

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

Comrie, Bernard. 2000. Language contact, lexical borrowing, and semantic fields. In Dicky Gilbers, John Nerbonne & Jos Schaeken (eds.), *Languages in Contact*, 73–86. Amsterdam-Atlanta, GA: Rodopi.

Cruz, Emiliana & Anthony C Woodbury. 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.

Dediu, Dan & Stephen C. Levinson. 2012. Abstract Profiles of Structural Stability Point to Universal Tendencies, Family-Specific Factors, and Ancient Connections between Languages. *PLoS ONE* 7(9).

DiCanio, Christian & Ryan Bennett. 2020. Mesoamerica. In Carlos Gussenhoven & Aoju Chen (eds.), *The Oxford Handbook of Language Prosody*, chap. 28, 408–427. Oxford University Press.

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

DiCanio, Christian T. 2022a. Itunyoso Triqui Collection of Christian DiCanio. The Archive of the Indigenous Languages of Latin America, ailla.utexas.org. PID ailla:243667; accessed January 5, 2022.

DiCanio, Christian T. 2022b. Vocabulario del Triqui de San Martín Itunyoso. unpublished.

Dimmendaal, Gerrit Jan. 2011. *Historical linguistics and the comparative study of African languages.* Amsterdam: John Benjamins.

Dockum, Rikker. 2019. *The tonal comparative method: Tai tone in historical perspective*: Yale University PhD dissertation.

Driver, Harold Edson & Alfred Louis Kroeber. 1932. Quantitative expression of cultural relationships. *University of California Publications in American Archaeology and Ethnology* 31(4). 211–256.

Drummond, Alexei J, Simon Y W Ho, Matthew J Phillips & Andrew Rambaut. 2006. Relaxed phylogenetics and dating with confidence. *PLoS biology* 4(5). e88.

Duchene, Sebastian, Philippe Lemey, Tanja Stadler, Simon YW Ho, David A Duchene, Vijaykrishna Dhanasekaran & Guy Baele. 2020. Bayesian evaluation of temporal signal in measurably evolving populations. *Molecular Biology and Evolution* 37(11). 3363–3379.

Dunn, Michael J., Simon J. Greenhill, Stephen C. Levinson & Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473. 79–82.

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.

Egland, Steven T. 1983. *La inteligibilidad interdialectal en México: Resultados de algunos sondeos.* México, D.F.: Instituto Lingüístico de Verano, A.C.

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.

Ferlus, Michel. 2004. The origin of tones in Viet-Muong. In Somsonge Buruspat (ed.), *Papers from the Eleventh Annual Meeting of the Southeast Asian Linguistics Society 2001*, 297–313. Tempe, Arizona: Arizona State University Programme for Southeast Asian Studies Monograph Series Press.

FitzJohn, Richard G, Wayne P Maddison & Sarah P Otto. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Systematic Biology* 58(6). 595–611.

Flannery, Kent V. & Joyce Marcus (eds.). 1983. *The cloud people: Divergent evolution of the Zapotec and Mixtec civilizations.* New York: Academic Press.

François, Alexandre. 2015. Trees, waves and linkages: Models of language diversification. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge Handbook of Historical Linguistics*, 161–189. Routledge.

Fritz, Susanne A & Andy Purvis. 2010. Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conservation Biology* 24(4). 1042–1051.

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

Gedney, William J. 1972. A checklist for determining tones in Tai dialects. In M. Estellie Smith (ed.), *Studies in Linguistics in Honor of George L. Trager*, 423–437. The Hague: Mouton.

Gerfen, Henry James. 1996. *Topics in the phonology and phonetics of Coatzospan Mixtec*: University of Arizona PhD dissertation.

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

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

Gray, Russell D., David Bryant & Simon J. Greenhill. 2010. On the shape and fabric of human history. *Philosophical Transactions of the Royal Society B* 365. 3923–3933.

Greenhill, Simon J., Thomas E Currie & Russell D. Gray. 2009. Does horizontal transmission invalidate cultural phylogenies? *Proceedings of the Royal Society of London B: Biological Sciences* 276. 2299–2306.

Greenhill, Simon J & Russell D Gray. 2009. Austronesian language phylogenies: Myths and misconceptions about Bayesian computational methods. In Alexander Adelaar & Andrew Pawley (eds.), *Austronesian historical linguistics and culture history: a festschrift for Robert Blust*, 375–397. Canberra: Pacific Linguistics.

Greenhill, Simon J, Paul Heggarty & Russell D Gray. 2020. Bayesian phylolinguistics. In Richard D. Janda, Brian D. Joseph & Barbara S. Vance (eds.), *The Handbook of Historical Linguistics*, vol. 2, chap. 11, 226–253. Wiley Blackwell.

Gudschinsky, Sarah C. 1958. Mazatec dialect history: a study in miniature. *Language* 34(4). 469–481.

Gudschinsky, Sarah C. 1959. *Proto-Popotecan: a comparative study of Popolocan and Mixtecan.* Baltimore: Waverly Press.

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

Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2022. Glottolog 4.6. Available online at `http://glottolog.org`.

Heggarty, Paul, Warren Maguire & April McMahon. 2010. Splits or waves? Trees or webs? How divergence measures and network analysis can unravel language histories. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365(1559). 3829–3843.

Hernández Martínez, Carmen & Sandra Auderset. 2020. *San Martín Duraznos Mixtec Vocabulary.* unpublished 1st edn.

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.

Hoffmann, Konstantin, Remco Bouckaert, Denise Kühnert & Simon J. Greenhill. 2021. Bayesian phylogenetic analysis of linguistic data using BEAST. *Journal of Language Evolution* 6(2). 119–135.

Hoijer, Harry. 1956. Lexicostatistics: a critique. *Language* 32(1). 49–60.

Holland, William R. 1959. Dialect variations of the Mixtec and Cuicatec areas of Oaxaca, Mexico. *Anthropological Linguistics* 1(8). 25–31.

Hollenbach, Barbara E. 1977. Phonetic vs. phonemic correspondence in two Trique dialects. In William R. Merrifield (ed.), *Studies in Otomanguean phonology*, Dallas: Summer Institute of Linguistics and the University of Texas at Arlington.

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 E. 2017. *Diccionario mixteco de Magdalena Peñasco.* Instituto Lingüístico de Verano, A.C.

Holman, Eric W. 2010. Do languages originate and become extinct at constant rates? *Diachronica* 27(2). 214–225.

Huson, Daniel H & David Bryant. 2006. Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution* 23(2). 254–267.

Hübler, Nataliia. 2022. Phylogenetic signal and rate of evolutionary change in language structures. *Royal Society Open Science* 9(211252).

INALI. 2009a. *Catálogo de las Lenguas Indígenas Nacionales.* México, D.F.: Instituto Nacional de Lenguas Indígenas INALI.

INALI. 2009b. mixteco. In *Catálogo de las Lenguas Indígenas Nacionales*, 199–218. México, D.F.: Instituto Nacional de Lenguas Indígenas INALI.

Jacques, Guillaume & Johann-Mattis List. 2019. Save the trees: Why we need tree models in linguistic reconstruction (and when we should apply them). *Journal of historical linguistics* 9(1). 128–167.

Janda, Richard D & Brian D Joseph. 2003. On language, change, and language change–or, of history, linguistics, and historical linguistics. In Brian D. Joseph & Richard D. Janda (eds.), *The Handbook of Historical Linguistics*, 3–180. Oxford: Blackwell.

Jansen, Maarten. 1990. The search for history in Mixtec codices. *Ancient Mesoamerica* 1. 99–112.

Jansen, Maarten & Gabina Aurora Pérez Jiménez. 2011. Introduction To Mixtec Pictography. In *The Mixtec Pictorial Manuscripts*, chap. 1, 1–41. Brill.

Jiménez Moreno, Wigberto. 1962. Estudios mixtecos. In *Vocabulario en lengua mixteca*, 9–105. Instituto Nacional Indigenista e Instituto Nacional de Antropología.

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.

Joseph, Umbavu V & Robbins Burling. 2001. Tone correspondences among the Boro languages. *Linguistics of the Tibeto-Burman Area* 24(2). 41–55.

Josserand, J. Kathryn. 1983. *Mixtec dialect history*: Tulane University PhD dissertation.

Josserand, J. Kathryn, Marcus Winter & Nicholas A Hopkins. 1984. *Essays in Otomanguean culture history.* Nashville, Tennessee: Vanderbilt University.

Joyce, Arthur A. 2011. *Mixtecs, Zapotecs, and Chatinos: ancient peoples of southern Mexico.* Wiley & Sons.

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

Kalyan, Siva & Alexandre François. 2018. Freeing the Comparative Method from the tree model. *Senri Ethnological Studies* 98. 59–89.

Kass, Robert E & Adrian E Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90(430). 773–795.

Kassian, Alexei, George Starostin, Anna Dybo & Vasiliy Chernov. 2010. The Swadesh wordlist. An attempt at semantic specification. *The Journal of Language Relationship* 4. 46–89.

Kaufman, Terrence. 1983. New Perspectives on Comparative Otomanguean Phonology. Unpublished monograph.

Kaufman, Terrence. 1988. Otomanguean tense/aspect/mood, voice, and nominalization markers. Unpublished monograph.

Kaufman, Terrence. 1990. Language history in South America: What we know and how to know more. In Doris Payne (ed.), *Amazonian linguistics: Studies in lowland South American languages*, 13–31. Austin: University of Texas Press.

Kaufman, Terrence. 2006. Oto-Manguean languages. In Keith Brown (ed.), *Encyclopedia of Language & Linguistics*, vol. 9, 118–124. Oxford: Elsevier.

Kaufman, Terrence. 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 Søren Wichmann (ed.), *Languages and linguistics of Mexico and Northern Central America: a comprehensive guide*, De Gruyter Mouton.

Kolipakam, Vishnupriya, Fiona M Jordan, Michael Dunn, Simon J Greenhill, Remco Bouckaert, Russell D Gray & Annemarie Verkerk. 2018. A Bayesian phylogenetic study of the Dravidian language family. *Royal Society Open Science* 5(3).

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.

Laycock, Donald C. 1970. Eliciting basic vocabulary in New Guinea. In *Pacific linguistic studies in honour of Arthur Capell*, 1127–1176. Canberra: Australian National University.

Lee, Ok Joo. 2022. Tones of Asian languages. In Chris Shei & Saihong Li (eds.), *The Routledge Handbook of Asian Linguistics*, chap. 14. Taylor & Francis.

Lee, Sean & Toshikazu Hasegawa. 2011. Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages. *Proceedings of the Royal Society B: Biological Sciences* 278(1725). 3662–3669.

Lewis, M. Paul, Gary F. Simons & Charles D. Fennig. 2015. *Ethnologue: Languages of the World*. Dallas: SIL International 18th edn.

List, Johann-Mattis, Simon Greenhill, Tiago Tresoldi & Robert Forkel. 2019. LingPy. A Python library for historical linguistics. Version 2.5.6. `http://lingpy.org`, DOI: https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy.

List, Johann-Mattis, Simon J Greenhill & Russell D Gray. 2017. The potential of automatic word comparison for historical linguistics. *PLoS ONE* 12(1). e0170046.

List, Johann-Mattis, Mary Walworth, Simon J Greenhill, Tiago Tresoldi & Robert Forkel. 2018. Sequence comparison in computational historical linguistics. *Journal of Language Evolution* 3(2). 130–144.

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.

Longacre, Robert E. 1962. Amplification of Gudschinsky's Proto-Popolocan-Mixtecan. *International Journal of American Linguistics* 28(4). 227–242.

Longacre, Robert E. 1966. On linguistic affinities of Amuzgo. *International Journal of American Linguistics* 32(1). 46–49.

Macaulay, Monica & Joseph C. Salmons. 1995. The phonology of glottalization in Mixtec. *International Journal of American Linguistics* 61(1). 38–61.

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

Macklin-Cordes, Jayden L, Claire Bowern & Erich R Round. 2021. Phylogenetic signal in phonotactics. *Diachronica* 38(2). 210–258.

Maddieson, Ian. 2013. Tone. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*, Leipzig: Max Planck Institute for Evolutionary Anthropology.

Maddison, David R, David L Swofford & Wayne P Maddison. 1997. NEXUS: an extensible file format for systematic information. *Systematic biology* 46(4). 590–621.

Maddison, Wayne P, Peter E Midford & Sarah P Otto. 2007. Estimating a binary character's effect on speciation and extinction. *Systematic Biology* 56(5). 701–710.

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

Martin, J N. 2020. Diccionario del Mixteco de El Jicaral y Otros Pueblos Fronterizos de Guerrero y Oaxaca. unpublished.

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

Matsukawa, Kosuke. 2008. Reconstruction of Proto-Trique phonemes. *University of Pennsylvania Working Papers in Linguistics* 14(1).

Maturana Russel, Patricio, Brendon J Brewer, Steffen Klaere & Remco R Bouckaert. 2019. Model selection and parameter inference in phylogenetics using nested sampling. *Systematic biology* 68(2). 219–233.

McKendry, Inga. 2013. *Tonal association, prominence and prosodic structure in South-Eastern Nochixtlán Mixtec*. Edinburgh: University of Edinburgh PhD dissertation.

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

Mendoza Ruíz, Juana. 2016. *Fonología segmental y patrones tonales del Tu'un Savi de Alcozauca de Guerrero*. México, D.F. CIESAS MA thesis.

Morey, Stephen. 2005. Tonal change in the Tai languages of Northeast India. *Linguistics of the Tibeto-Burman Area* 28(2). 139–202.

Münkemüller, Tamara, Sébastien Lavergne, Bruno Bzeznik, Stéphane Dray, Thibaut Jombart, Katja Schiffers & Wilfried Thuiller. 2012. How to measure and test phylogenetic signal. *Methods in Ecology and Evolution* 3(4). 743–756.

Nettle, Daniel. 1999. Is the rate of linguistic change constant? *Lingua* 108(2-3). 119–136.

Nichols, Johanna. 2003. Diversity and stability in language. In Richard D. Janda & Brian D. Joseph (eds.), *Handbook of Historical Linguistics*, 283–310. London: Blackwell.

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.

Orme, David, Gavin Thomas Freckleton, Thomas Petzold, Susanne Fritz, Nick Isaac & Will Pearse. 2018. *caper: Comparative Analyses of Phylogenetics and Evolution in R* r package version 1.0.1 edn. `https://CRAN.R-project.org/package=caper`.

Padgett, Erin Padgett. 2017. *Tools For Assessing Relatedness In Understudied Language Varieties: A Survey Of Mixtec Varieties In Western Oaxaca, Mexico*. Grand Forks, ND University of North Dakota MA thesis.

Pankratz, Leo & Eunice V Pike. 1967. Phonology and morphotonemics of Ayutla Mixtec. *International*

Journal of American Linguistics 33(4). 287–299.

Penny, David, Bennet J McComish, Michael A Charleston & Michael D Hendy. 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *Journal of Molecular Evolution* 53(6). 711–723.

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.

Peters, Simon L. 2018. *The inventory and distribution of tone in Tù'un Ndá'vi, the Mixtec of Piedra Azul (San Martín Peras), Oaxaca.* University of California Santa Barbara MA thesis.

Phillips, Joshua & Claire Bowern. 2022. Bayesian methods for ancestral state reconstruction in morphosyntax: Exploring the history of argument marking strategies in a large language family. *Journal of Language Evolution* .

Pike, Eunice V & John H Cowan. 1967. Huajuapan Mixtec phonology and morphophonemics. *Anthropological Linguistics* 1–15.

Pike, Kenneth L. 1948. Tonemic perturbations in Mixteco, with special emphasis on tonomechanical subclasses. In *Tone languages*, 77–94. Ann Arbor: University of Michigan Press.

Pérez Rodríguez, Verónica. 2013. Recent Advances in Mixtec Archaeology. *Journal of Archaeological Research* 21(1). 75–121.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. `https://www.R-project.org/`.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. `https://www.R-project.org/`.

Rambaut, Andrew, Alexei J. Drummond, Dong Xie, Guy Baele & Marc A. Suchard. 2018. Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology* 67(5). 901–904.

Ramirez, Yesica. 2020. Vocabulario del Mixteco de San Juan Mixtepec. unpublished.

Rankin, Robert L. 2003. The comparative method. In Brian D. Joseph & Richard D. Janda (eds.), *The Handbook of Historical Linguistics*, chap. 1, 181–212. Wiley Blackwell.

Ratliff, Martha. 2015. Tonoexodus, tonogenesis, and tone change. In Patrick Honeybone & Joseph Salmons (eds.), *The Oxford Handbook of Historical Phonology*, 245–261. Oxford: Oxford University Press.

Rensch, Calvin Ross. 1976. *Comparative Otomanguean Phonology.* Bloomington: Indiana University Press.

Revell, Liam J, Luke J Harmon & David C Collar. 2008. Phylogenetic signal, evolutionary process, and rate. *Systematic Biology* 57(4). 591–601.

de los Reyes, Antonio. 1890 [1593]. *Arte en lengua mixteca.* Alençon: Typographie E. Renaut de Broise.

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

Ross, Malcolm. 1988. *Proto Oceanic and the Austronesian languages of western Melanesia.* Australian National University.

Ross, Malcolm & Mark Durie. 1996. Introduction. In Mark Durie & Malcolm Ross (eds.), *The comparative method reviewed: Regularity and irregularity in language change*, chap. 1, 3–38. Oxford University Press.

Ross, Malcolm D. 1996. Contact-induced change and the comparative method: cases from Papua New Guinea. In Mark Durie & Malcolm D. Ross (eds.), *The comparative method reviewed: regularity and irregularity in language change*, 180 – 217. New York: Oxford University Press.

Salazar, Jeremías, Saleem Alfaife, Guillem Belmar Viernes, Eric W. Campbell, Gabriel Mendoza, Olguín-Martínez Jesús, Griselda Reyes Basurto, Catherine Scanlon, Giorgia Troiani & Alonso Vásquez-Aguilar. 2021. Vocabulario de Yucunani Sà'án Sàvǐ (Mixteco). unpublished.

Schmidt, Johannes. 1872. *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen.* Weimar: H. Böhlau.

Schuchardt, Hugo. 1900. *Über die Klassifikation der romanischen Mundarten. Probevorlesung gehalten zu Leipzig am 30. April 1870.* Graz: Styria.

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.

Smith Stark, Thomas C. 1994. El estado actual de los estudios de las lenguas Mixtecanas y Zapotecanas. In Leonardo Manrique, Yolanda Lastra & Doris Bartholomew (eds.), *Panorama de los estudios de las lenguas indígenas de México: Tomo II*, vol. 17, 5–186. Quito: Abya-Yala.

Smith-Stark, Thomas C. 1994. Mesoamerican calques. In Carolyn J. MacKay & Verónica Vázquez (eds.), *Investigaciones lingüísticas en Mesoamérica*, 15–50. México, D.F.: Universidad Nacional Autónoma de México.

Solano, Juvenal. 2020. Vocabulario del Mixteco de San Sebastián del Monte. unpublished.

Stacklies, Wolfram, Henning Redestig, Matthias Scholz, Dirk Walther & Joachim Selbig. 2007. pcaMethods – a Bioconductor package providing PCA methods for incomplete data. *Bioinformatics* 23(9). 1164–1167.

Stadler, Tanja, Denise Kühnert, Sebastian Bonhoeffer & Alexei J Drummond. 2013. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences* 110(1). 228–233.

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.

Swadesh, Morris. 1967. Lexicostatistic classification. In Norman A McQuown (ed.), *Handbook of Middle American Indians*, vol. 5: Linguistics, 79–115. Austin: University of Texas Press.

Swanton, Michael & Juana Mendoza Ruíz. 2021. Observaciones sobre la diacronía del tono en el Tu'un Savi (mixteco) de Alcozauca de Guerrero. In Francisco Arellanes & Lilián 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.

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

Vázquez, Octavio León. 2017. *Sandhi tonal en el mixteco de Yucuquimi de Ocampo, Oaxaca.* México, D.F. CIESAS MA thesis.

Willems, Matthieu, Etienne Lord, Louise Laforest, Gilbert Labelle, François-Joseph Lapointe, Anna Maria Di Sciullo & Vladimir Makarenkov. 2016. Using hybridization networks to retrace the evolution of Indo-European languages. *BMC Evolutionary Biology* 16(180).

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.

Witzlack-Makarevich, Alena, Johanna Nichols, Kristine Hildebrandt, Taras Zakharko & Balthasar Bickel. 2022. Managing AUTOTYP data: Design principles and implementation. In Eve Koller Lauren B. Collister Andrea L. Berez-Kroeker, Bradley McDonnell (ed.), *The Open Handbook of Linguistic Data Management*, Cambridge: MIT Press.

Yip, Moira. 2002. *Tone.* Cambridge University Press.

Yu, Guangchuang, David K Smith, Huachen Zhu & Tommy Tsan-Yuk Lam. 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* 8(1). 28–36.

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.

Appendix A

Appendix to Chapter 1

Table A.1: Varieties and sources (varieties marked with an asterisk were excluded from the final analysis due to low coverage)

No.	Variety	Branch	Sources
1	ConcepcionPapaloCuicatec*	Cuicatec	Bradley 1991
2	SanJuanTepeuxila1900Cuicatec	Cuicatec	Belmar 1902
3	SantaMariaPapaloCuicatec	Cuicatec	Anderson & Roque 1983
4	AbasoloValleMixtec	Mixtec	Galindo Sánchez 2009
5	AlacatlatzalaMixtec	Mixtec	Zylstra 2012; Anderson 2006; Zylstra 1991; Josserand 1983; Dürr 1987
6	AlcozaucaGuerreroMixtec	Mixtec	Swanton & Mendoza Ruíz 2021; Josserand 1983
7	CahuatacheMixtec	Mixtec	Dürr 1987; Josserand 1983
8	ChalcatongoHidalgoMixtec	Mixtec	Swanton & Mendoza Ruíz 2021; Macaulay 1996; Josserand 1983
9	CoicoyanlasFloresMixtec	Mixtec	Josserand 1983
10	CosoltepecMixtec	Mixtec	Josserand 1983
11	CuatzoquitengoMixtec	Mixtec	Josserand 1983
12	CuilapamGuerreroMixtec*	Mixtec	Josserand 1983
13	CuyamecalcoVillaZaragozaMixtec	Mixtec	Josserand 1983
14	ElJicaralMixtec	Mixtec	Martin 2020
15	ElRosarioMicaltepecMixtec*	Mixtec	Josserand 1983
16	GuadalupeVillahermosaMixtec	Mixtec	Josserand 1983
17	IxpantepecNievesMixtec	Mixtec	Josserand 1983
18	LaBateaMixtec	Mixtec	Ramirez 2020
19	LosTejocotesMixtec*	Mixtec	Josserand 1983
20	MagdalenaPenascoMixtec	Mixtec	Hollenbach 2017

21	MetlatonocMixtec*	Mixtec	Dürr 1987; Josserand 1983
22	PiedraAzulMixtec	Mixtec	Mendoza & Peters 2020
23	PinotepaDonLuisMixtec	Mixtec	Josserand 1983
24	SanAgustinAtenangoMixtec	Mixtec	Josserand 1983
25	SanAgustinChayucoMixtec	Mixtec	Dürr 1987; Josserand 1983; Pensinger et al. 1974
26	SanAgustinMonteLobosMixtec*	Mixtec	Josserand 1983
27	SanAgustinTlacotepecMixtec	Mixtec	Josserand 1983
28	SanAndresNuxinoMixtec	Mixtec	Josserand 1983
29	SanAndresYutatioMixtec	Mixtec	Williams et al. 2017
30	SanAntonioHuitepecMixtec	Mixtec	Josserand 1983
31	SanAntonioNduayacoMixtec*	Mixtec	Josserand 1983
32	SanAntonioTepetlapaMixtec	Mixtec	Josserand 1983
33	SanBartolomeYucuaneMixtec	Mixtec	Josserand 1983
34	SanBartoloSoyaltepecMixtec	Mixtec	Josserand 1983
35	SanCristobalJamiltepecMixtec*	Mixtec	Josserand 1983
36	SanEstebanAtatlahucaMixtec	Mixtec	Dürr 1987; Josserand 1983; Alexander 1980
37	SanFranciscoJaltepetongoMixtec*	Mixtec	Josserand 1983
38	SanFranciscolasFloresMixtec*	Mixtec	Josserand 1983
39	SanFranciscoSayultepecMixtec	Mixtec	Josserand 1983
40	SanJeronimoProgresoMixtec	Mixtec	Shields 1988; Dürr 1987; Josserand 1983
41	SanJeronimoXayacatlanMixtec	Mixtec	Dürr 1987; Josserand 1983
42	SanJorgeNuchitaMixtec	Mixtec	Josserand 1983
43	SanJoseSinicahuaMixtec*	Mixtec	Josserand 1983
44	SanJuanCoatzospamMixtec	Mixtec	Small 1990; Dürr 1987; Josserand 1983
45	SanJuanColoradoMixtec	Mixtec	Stark et al. 1986; Josserand 1983
46	SanJuanDiuxiMixtec	Mixtec	Kuiper & Oram 1991; Dürr 1987; Josserand 1983
47	SanJuanMixtepecMixtec	Mixtec	Dürr 1987; Josserand 1983;
48	SanJuanNumiMixtec	Mixtec	Josserand 1983
49	SanJuanTamazolaMixtec	Mixtec	Josserand 1983
50	SanJuanTeitaMixtec	Mixtec	Josserand 1983
51	SanJuanYutaMixtec*	Mixtec	Josserand 1983
52	SanLorenzoMixtec	Mixtec	Josserand 1983
53	SanLuisMoreliaMixtec	Mixtec	Josserand 1983
54	SanMarcoslaFlorMixtec	Mixtec	Anonymous p.c.
55	SanMartinDuraznosMixtec	Mixtec	Hernández Martínez & Auderset 2020; Josserand 1983
56	SanMartinEstadoMixtec	Mixtec	Josserand 1983

57	SanMartinPerasMixtec	Mixtec	Josserand 1983
58	SanMateoPenascoMixtec*	Mixtec	Josserand 1983
59	SanMateoSindihuiMixtec	Mixtec	Josserand 1983
60	SanMiguelAchiutlaMixtec	Mixtec	Josserand 1983
61	SanMiguelAhuehuetitlanMixtec	Mixtec	Josserand 1983
62	SanMiguelAmatitlanMixtec*	Mixtec	Josserand 1983
63	SanMiguelChicahuaMixtec	Mixtec	Josserand 1983
64	SanMiguelElGrandeMixtec	Mixtec	Dürr 1987; Josserand 1983
65	SanMiguelIxtapamMixtec*	Mixtec	Josserand 1983
66	SanMiguelPiedrasMixtec	Mixtec	Josserand 1983
67	SanMiguelProgresoMixtec*	Mixtec	Josserand 1983
68	SanMiguelTlacotepecMixtec	Mixtec	Josserand 1983
69	SanPedroAtoyacMixtec	Mixtec	Josserand 1983
70	SanPedroChayucoMixtec*	Mixtec	Josserand 1983
71	SanPedroCoxcaltepecCantarosMixtec*	Mixtec	Josserand 1983
72	SanPedroJicayanMixtec	Mixtec	Josserand 1983
73	SanPedroJocotipacMixtec	Mixtec	Josserand 1983
74	SanPedroMolinosMixtec	Mixtec	Dürr 1987; Josserand 1983
75	SanPedroTidaaMixtec	Mixtec	Josserand 1983
76	SanPedroTututepecMixtec	Mixtec	Josserand 1983
77	SanPedroYosonamaMixtec	Mixtec	Gittlen 2016
78	SanPedroySanPablo Teposcolula1600Mixtec	Mixtec	Josserand 1983; de Alvarado 1962 [1593]
79	SanSebastianMonteMixtec	Mixtec	Solano 2020; Josserand 1983
80	SanSebastianTecomaxtlahuacaMixtec	Mixtec	Josserand 1983
81	SantaAnaCuauhtemocMixtec	Mixtec	Josserand 1983
82	SantaCatarinaAdequezMixtec*	Mixtec	Josserand 1983
83	SantaCatarinaEstetlaMixtec	Mixtec	Josserand 1983
84	SantaCatarinaMechoacanMixtec	Mixtec	Josserand 1983
85	SantaCatarinaTlaltempanMixtec	Mixtec	Josserand 1983
86	SantaCruzBravoMixtec	Mixtec	Josserand 1983
87	SantaCruzItundujiaMixtec	Mixtec	Josserand 1983
88	SantaCruzNundacoMixtec*	Mixtec	Josserand 1983
89	SantaLuciaMonteverdeMixtec	Mixtec	Josserand 1983
90	SantaMariaAcatepecMixtec	Mixtec	Josserand 1983
91	SantaMariaApazcoMixtec	Mixtec	Josserand 1983
92	SantaMariaChigmecatitlanMixtec	Mixtec	Josserand 1983
93	SantaMariaHuazolotitlanMixtec	Mixtec	Josserand 1983
94	SantaMariaJicaltepecMixtec	Mixtec	Dürr 1987; Josserand 1983
95	SantaMariaNutioMixtec*	Mixtec	Josserand 1983
96	SantaMariaPenolesMixtec	Mixtec	Dürr 1987; Josserand 1983
97	SantaMariaTataltepecMixtec*	Mixtec	Josserand 1983

98	SantaMariaYolotepecMixtec	Mixtec	Josserand 1983
99	SantaMariaYucuhitiMixtec	Mixtec	Josserand 1983
100	SantaMariaYucunicocoMixtec	Mixtec	Josserand 1983
101	SantaMariaZacatepecMixtec	Mixtec	Swanton & Mendoza Ruíz 2021; Towne 2011; Josserand 1983
102	SantiagoApoalaMixtec	Mixtec	Josserand 1983
103	SantiagoCacaloxtepecMixtec	Mixtec	Dürr 1987; Josserand 1983
104	SantiagoChazumbaMixtec	Mixtec	Josserand 1983
105	SantiagoIxtaltepecMixtec	Mixtec	Josserand 1983
106	SantiagoIxtayutlaMixtec	Mixtec	Josserand 1983
107	SantiagoJamiltepecMixtec	Mixtec	Johnson 1988; Josserand 1983
108	SantiagoJuxtlahuacaMixtec	Mixtec	Josserand 1983
109	SantiagoNundicheMixtec*	Mixtec	Josserand 1983
110	SantiagoNuyooMixtec	Mixtec	Josserand 1983
111	SantiagoPinotepaNacionalMixtec	Mixtec	Josserand 1983
112	SantiagoTamazolaMixtec	Mixtec	Josserand 1983
113	SantiagoTilantongoMixtec	Mixtec	Josserand 1983
114	SantiagoTlazoyaltepecMixtec	Mixtec	Josserand 1983
115	SantiagoYosonduaMixtec	Mixtec	Farris 1992; Josserand 1983
116	SantoDomingoHuendioMixtec	Mixtec	Becerra Roldán 2015
117	SantoDomingoNundoMixtec*	Mixtec	Josserand 1983
118	SantoDomingoNuxaaMixtec	Mixtec	Josserand 1983
119	SantoDomingoTonahuixtlaMixtec	Mixtec	Josserand 1983
120	SantosReyesTepejilloMixtec	Mixtec	Josserand 1983
121	SantoTomasOcotepecMixtec	Mixtec	Alexander 1988; Dürr 1987; Josserand 1983
122	TepangoMixtec	Mixtec	Hills 1990; Dürr 1987; Josserand 1983
123	TepejilloMixtec	Mixtec	Josserand 1983
124	TlahuapaMixtec	Mixtec	Reyes Basurto 2020
125	TotoltepecGuerreroMixtec*	Mixtec	Josserand 1983
126	XayacatlanBravoMixtec	Mixtec	Josserand 1983
127	XochapaMixtec	Mixtec	Stark et al. 2013
128	YoloxochitlMixtec	Mixtec	Amith & Castillo García n.d.; Josserand 1983
129	YucunaniMixtec	Mixtec	Salazar et al. 2021
130	YucunutiBenitoJuarezMixtec	Mixtec	Josserand 1983
131	YucuquimiOcampoMixtec	Mixtec	Swanton & Mendoza Ruíz 2021; Josserand 1983
132	YutanduchiGuerreroMixtec*	Mixtec	Josserand 1983
133	ZapotitlanPalmasMixtec	Mixtec	Josserand 1983
134	SanAndresChicahuaxtla1890Triqui	Triqui	Belmar 1897

135	SanAndresChicahuaxtlaTriqui	Triqui	Hernández Mendoza 2020; Good 1978
136	SanJuanCopalaTriqui	Triqui	Hollenbach 1992
137	SanMartinItunyosoTriqui	Triqui	DiCanio 2022b

Table A.2: Mixtec groups

Mixtec Variety	New Group	Josserand Area
San Pedro y San Pablo Teposcolula 1600	Unclear	Eastern Alta
Cuyamecalco Villa de Zaragoza	Group 1	Northern Alta
San Juan Coatzóspam	Group 1	Northern Alta
Santa Ana Cuauhtémoc	Group 1	Northern Alta
Santiago Ixtayutla	Group 2.1	Coast
Pinotepa de Don Luis	Group 2.2	Coast
San Antonio Tepetlapa	Group 2.2	Coast
San Francisco Sayultepec	Group 2.2	Coast
San Juan Colorado	Group 2.2	Coast
San Pedro Atoyac	Group 2.2	Coast
San Pedro Jicayán	Group 2.2	Coast
Santa María Jicaltepec	Group 2.2	Coast
Santa María Zacatepec	Group 2.2	Coast
San Agustín Chayuco	Group 2.2.1	Coast
San Lorenzo	Group 2.2.1	Coast
San Pedro Tututepec	Group 2.2.1	Coast
Santa Catarina Mechoacán	Group 2.2.1	Coast
Santa María Acatepec	Group 2.2.1	Coast
Santa María Huazolotitlán	Group 2.2.1	Coast
Santiago Jamiltepec	Group 2.2.1	Coast
Santiago Pinotepa Nacional	Group 2.2.1	Coast
San Bartolomé Yucuañe	Group 3.1	Western Alta
San Juan Teita	Group 3.1	Western Alta
Magdalena Peñasco	Group 3.2	Western Alta
San Agustín Tlacotepec	Group 3.2	Western Alta
San Miguel Achiutla	Group 3.2	Western Alta
Santa María Yolotepec	Group 3.2	Western Alta
Santo Domingo Huendio	Group 3.2	Western Alta
San Antonio Huitepec	Group 4	Eastern Alta
San Juan Tamazola	Group 4	Eastern Alta
San Mateo Sindihui	Group 4	Eastern Alta
San Miguel Piedras	Group 4	Eastern Alta
San Andrés Nuxiño	Group 4.1	Eastern Alta

Santo Domingo Nuxaá	Group 4.1	Eastern Alta
Santa Catarina Estetla	Group 4.2	Eastern Alta
Santa María Peñoles	Group 4.2	Eastern Alta
Santiago Tlazoyaltepec	Group 4.2	Eastern Alta
San Juan Diuxi	Linkage 5	Eastern Alta
San Pedro Tidaá	Linkage 5	Eastern Alta
Santiago Tilantongo	Linkage 5	Eastern Alta
San Juan Ñumi	Linkage 5	Western Alta
San Pedro Yosoñama	Linkage 5	Western Alta
San Bartolo Soyaltepec	Group 5.1	Northeastern Alta
San Miguel Chicahua	Group 5.1	Northeastern Alta
San Pedro Jocotipac	Group 5.1	Northeastern Alta
Santa María Apazco	Group 5.1	Northeastern Alta
Santiago Apoala	Group 5.1	Northeastern Alta
Santiago Ixtaltepec	Group 5.1	Northeastern Alta
Santo Tomás Ocotepec	Group 6.1	Western Alta
Santa María Yucuhiti	Group 6.2.1	Western Alta
Santiago Nuyoó	Group 6.2.1	Western Alta
Chalcatongo de Hidalgo	Group 6.2.2	Western Alta
San Esteban Atatlahuca	Group 6.2.2	Western Alta
San Miguel El Grande	Group 6.2.2	Western Alta
San Pedro Molinos	Group 6.2.2	Western Alta
Santa Cruz Itundujia	Group 6.2.2	Western Alta
Santa Lucía Monteverde	Group 6.2.2	Western Alta
Santiago Yosondúa	Group 6.2.2	Western Alta
Santos Reyes Tepejillo	Group 7	Southern Baja
Abasolo del Valle	Group 7.1	Mixtepec
La Batea	Group 7.1	Mixtepec
San Juan Mixtepec	Group 7.1	Mixtepec
Santa María Yucunicoco	Group 7.1	Mixtepec
Yucunani	Group 7.1	Mixtepec
Santiago Juxtlahuaca	Group 7.1	Southern Baja
Ixpantepec Nieves	Group 7.2	Southern Baja
San Martín Duraznos	Group 7.2	Southern Baja
San Sebastián Tecomaxtlahuaca	Group 7.2	Southern Baja
Coicoyán de las Flores	Group 7.3	Southern Baja
El Jicaral	Group 7.3	Southern Baja
Piedra Azul	Group 7.3	Southern Baja
San Jerónimo Progreso	Group 7.3	Southern Baja
San Marcos de la Flor	Group 7.3	Southern Baja
San Martín Peras	Group 7.3	Southern Baja
Alacatlatzala	Group 7.3.1	Guerrero

Alcozáuca de Guerrero	Group 7.3.1	Guerrero
Cahuatache	Group 7.3.1	Guerrero
Cuatzoquitengo	Group 7.3.1	Guerrero
Santa Cruz de Bravo	Group 7.3.1	Guerrero
Tlahuapa	Group 7.3.1	Guerrero
Xochapa	Group 7.3.1	Guerrero
Yoloxochitl	Group 7.3.1	Guerrero
Tepango	Group 7.3.1	Southern Baja
Cosoltepec	Group 7.4	Northern Baja
Santiago Chazumba	Group 7.4	Northern Baja
Santo Domingo Tonahuixtla	Group 7.4	Northern Baja
Tepejillo	Group 7.4	Northern Baja
Xayacatlán de Bravo	Group 7.4	Northern Baja
Zapotitlán Palmas	Group 7.4	Northern Baja
San Jerónimo Xayacatlán	Group 7.4.1	Northern Baja
Santa Catarina Tlaltempan	Group 7.4.1	Northern Baja
Santa María Chigmecatitlán	Group 7.4.1	Northern Baja
San Andrés Yutatío	Linkage 7.5	Tezoatlan
Santiago Cacaloxtepec	Linkage 7.5	Tezoatlan
Yucuñuti de Benito Juárez	Linkage 7.5	Tezoatlan
Yucuquimi de Ocampo	Linkage 7.5	Tezoatlan
Guadalupe Villahermosa (El Portesuelo)	Group 7.6	Central Baja
San Agustín Atenango	Group 7.6	Central Baja
San Jorge Nuchita	Group 7.6	Central Baja
San Luis Morelia	Group 7.6	Central Baja
San Sebastián del Monte	Group 7.6	Central Baja
San Martín del Estado	Group 7.6	Western Baja
San Miguel Ahuehuetitlán	Group 7.6	Western Baja
Santiago Tamazola	Group 7.6	Western Baja