Comprehensive database of sound changes reveals tree-like and wave-like processes in the Mixtec language family

The comparative method of historical linguistics based on the principle of regular sound change remains the primary tool for reconstructing proto-languages, describing sound changes, and understanding language relationships and linguistic history. However, the methods face challenges in situations where continued contact between languages produces so-called 'dialect areas' or 'dialect continua' (Ross & Durie 1996, Kalyan & François 2018). The Mixtec languages have been characterized as a dialect continuum (Longacre 1957, Josserand 1983). 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 US. Our work is based on an extensive database of Mixtec cognate sets, builds on previous work on proto-Mixtec reconstruction (Rensch 1976, Longacre 1957, Bradley & Josserand 1982, Josserand 1983, Dürr 1987, Kaufman in press), and includes novel data from ongoing documentation projects in the Mixteca.

We identify 245 segmental sound changes across 105 Mixtec varieties and analyze their distribution and conditioning environments (see Table 1). We investigate whether and to what extent the patterns and distribution of these sound changes align with previous classifications of the language family (Josserand 1983, Auderset et al. 2023). We build our inventory of sound changes in a bottom-up fashion, adding new changes as identified in the data. We establish the sound changes by evaluating each cognate set in light of the modern reflexes using the comparative method. To allow for the database to be expanded with more data in the future, we largely refrain from specifying environments with sound classes and rather list each conditioning environment separately. 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).

This fine-grained, bottom-up coding of the sound changes served as the basis for applying correlation measures, clustering techniques and principal components analysis (PCA). The former show that the distribution of most sound changes aligns well with previously proposed dialect areas (Josserand 1983) and subgroups (Auderset et al. 2023). In addition, the results of the PCA show that there are more tree-like and more wave-like parts of the family, rather than the whole family being characterized as a dialect continuum with overlapping distributions of sound changes. Our study adds detail to understanding Mixtec language history, which is of importance to communities and their language-related needs and goals in language maintenance and access. Our methods provide a model for handling large data sets of diversified language groups for refining questions in reconstruction and subgrouping that continued contact may have obscured.