TEI and the Mixtepec-Mixtec corpus: data integration, annotation and normalization of heterogeneous data for an under-resourced language

Jack Bowers [1,2,3]
Laurent Romary [2,4]
[1] Austrian Academy of Sciences - Austrian Center for Digital Humanities
[2] Inria - Team ALMAnaCH
[3] EPHE, Paris
[4]BBAW

Creating, editing and curating a corpus for an under-resourced language entails a number of unique demands, and in doing so, it is fundamentally essential that extra precautions are taken to ensure that what is created is extensible and reusable both by the language community and potential future researchers. This paper discusses our corpus creation using the TEI (Text Encoding Initiative, www.tei-c.org), as part of an ongoing language documentation project concerning the Mixtepec-Mixtec language (iso 639-3: mix)[1]. While the use of TEI for an indigenous language has been discussed with regards to digital dictionaries (Czaykowska-Higgins et al., 2014; Bowers and Romary, 2018), other than Bowers and Romary (2017) which gave a broad introductory overview of this project, there are no previous publications focusing on its use in addressing the specific needs of creating a corpus in the context of a language documentation project, or an under-resourced language.

When carrying out such a task it is essential to make the most out of every language resource available. In creating our multimodal corpus we integrate literally every source of the language we encounter, including: recorded spoken language (from consultation sessions and natural speech); time aligned speech transcriptions (exported from Praat annotations); a few dozen monolingual and bilingual booklets; original written materials created with project's speaker consultants; as well as less traditional sources such as: data from the small number of academic papers on the language; public safety pamphlets from the Mexican government; content from emails, text messages, social media posts, etc.

However, organizing and annotating an array of such heterogeneous sources within a common markup and data management system from which we can effectively search and extract key information presents a number of challenges with regards to: (a) the data representation formats

[1] Mixtepec-Mixtec (*Sa'an Savi* : 'rain language') is an Otomonguean spoken by roughly 9000-10000 people in the Juxtlahuaca district of Oaxaca, and parts of the Guerrero and Puebla states of Mexico. Most of the spoken data collected in this project originate from consultation sessions with two native speakers from a town called Yucunani (17.30083, -97.89389), which is part of the municipality of San Juan de Mixtepec.

(integrating metadata and a variety of diversely structured text files, including time aligned speech annotations); (b) multi-layered linguistic annotations and translations; (c) normalization of extensive variation in orthography (which is still under-development), and phonetic transcriptions from external sources.

Within the context of presenting the details of the issues listed above, we articulate our approaches to addressing them using TEI, and demonstrate how their application facilitates and improves search and extraction from the corpus using scripting and query database toolkits. Many topics covered and our solutions to them are not only relevant to Mixtepec-Mixtec, but to other under-resourced languages as well.

Works Cited

Bowers, J., & Romary, L. (2017). Language Documentation and Standards in Digital Humanities: TEI and the Documentation of Mixtepec-Mixtec. In A. Kawase (Ed.), *Proceedings of the 7th Conference of Japanese Association for Digital Humanities* (pp. 21–23). Kyoto, Japan: Doshisha University.

Bowers, J., & Romary, L. (2018). Bridging the gaps between digital humanities, lexicography and linguistics: a TEI dictionary for the documentation of Mixtepec-Mixtec. *Dictionaries: Journal of the Dictionary Society of North America*, *39*(2).

Czaykowska-Higgins, E., Holmes, M. D., & Kell, S. M. (2014). Using TEI for an Endangered Language Lexical Resource: The Nxaʔamxcín Database-Dictionary Project. *Language Documentation and Conservation*, *8*, 1–37.

TEI Consortium. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 3.4.0. July 2018. TEI Consortium. http://www.tei-c.org/Guidelines/P5/ (12/08/2018)