**Good**

**Lillehaugen**

**Ticha: A Digital text explorer for Colonial Zapotec**

The PI has assembled an excellent team of collaborators for the task at hand, representing a combination of expertise in Zapotec lexicography, morphology, syntax, and philology as well as in mark-up in TEI. Although far beyond Nahuatl in terms of the number of colonial texts, Central Valley Zapotec (CVZ) material does exist: most notably the Códoba's dictionary and grammar, well known to Zapotec scholars, and the bilingual Doctrina Christiana. In addition there are colonial documents which, according to the PI, comprise approximately 200 native-speaker documents, mostly wills and mostly in central valley Zapotec.

This Level 1 project has two basic inputs: (1) the Arte; and (2) the Doctrina. The former will be marked up in accord with TEI as demonstrated in Appendix C. The latter will be parsed in FLEx as demonstrated in Appendix B. CVZ material in the Arte will apparently be parsed and glossed manually (Appendix D) though perhaps in FLEx as well.

As the PI notes, the Arte is available online in an excellent digital facsimile through the John Carter Brown Library (<https://archive.org/details/arteenlenguazapo00juan>). The text is available in multiple formats (PDF, EPUB, Kindle, Daisy, Full Text, DiVu). Attempts were made to make the PDF searchable through OCR but this reviewer has found the JCB Library result to be very deficient in this regard as is, for the same reason of deficient OCR, the text rendering [Full Text]. It seems that the PI and team will not use the OCR-based Full Text from the John Carter Brown Library but, instead, transcribe the original text and then encode with TEI based XML elements and attributes.

The first major resource that this project will create is the TEI mark-up of the Arte (the extracted and interlinearized examples will be dealt with below). A sample of the mark-up is presented in Appendix C of the proposal. Students will "do a first pass translation of the Spanish portions of texts into modern Spanish and English, general editing, and proofreading of both English and Spanish. Others will also be employed to do first pass encoding in TEI" Valdivia will contribute 40 pages of encoding and review and edit TEI encoding for another 240 pages (apparently by the student assistants). I am slightly concerned about quality control of the conversion to modern Spanish and, more particularly, translation to English. I am also unsure of the necessity of translation to English as I think almost anyone wishing to study Zapotec would (should) have a good if not working knowledge of Spanish. The project is also not clear how many versions (original, modern Spanish, English) will be marked up in TEI. Finally, although a marked up text file is "searchable" in the broad sense, TEI does not in and of itself create an index, which would be extremely useful.

The second major resource that this project will create is an interlinearized and annotated version of the Doctrina (Appendix (which is already online at the JCB Library, with the same problems of searchability). This is an important and interesting project; it would provide a virtually unique resource on the morphosyntax and lexicon of colonial valley Zapotec. I would question, however, whether a project of this nature will have the impact on the humanities and the possibility of growth that would justify a Digital Humanities grant (as this proposal is for a "start-up"). An online parsed version of the Doctrina might be more appropriate for a Preservation and Access grant or other NEH initiatives that focus on a single resource. This is an important project that will hopefully receive NEH funding either through a start-up grant or another program. Nevertheless, as it stands now I think that there are several methodological issues that could benefit from greater clarification. Perhaps also, as explained below, a slight shift in focus would be beneficial.

According to my understanding FLEx, the program that will be used for the interlinearization of the Zapotec material in the grammar and the Zapotec side of the bilingual doctrina, applies a lexical database to a corpus to find matches. In the absence of a straight-up string match with basic roots/stems and bound morphemes (e.g., 'inactive' > 'in + active'), morphological parsers can work in several ways: (1) a set of rules (n > m / \_\_ p, enabling 'impossible' to be recognized as 'in + possible') or (2) a list of allomorphs with no attempt at generalization (e.g., /im/il/ir as allomorphs of /in/ ). The point of this observation is that the parsing of the Zapotec texts (the Doctrina and selected examples from the Arte) requires a lexical database. It is not clear from the proposal where this would originate as Cordoba's dictionary would probably not provide all the information (e.g., parsing out the irrealis or other aspect markers) necessary. Indeed, in reading this proposal it occurred to me that a dialectic relationship between the building of a lexicon capable of parsing the Doctrina and the testing of this lexicon against the Doctrina as it is interlinearized through FLEx would provide a unique way to demonstrate how progressive attempts at parsing a colonial text reveal lexicographic and grammatical information that might otherwise remain hidden and that could provide material for a more complete lexicon of CVZ. In essence, then, the dialect between a colonial Zapotec text and a lexical database (one constantly tested against the other through iterative interlinearization) would create both an parsed-glossed text and an enhanced dictionary.

The project also notes that the FLEx generated XML will be converted to TEI through XSLT. I am not sure if this is adequate. That is, an XML export from FLEx is quite simple. Each "word" is included in a hierarchical "group" where the input word is associated with various forms (e.g., morpheme/, gloss/, part.of.speech/. Converting to TEI of this alone is relatively simple, but to mark-up the Doctrina in TEI a separate level of encoding is needed and this must be done as it would be for the Arte and not simply through XSLT conversion of the FLEx output.

A caveat worth mentioning is that although the 200 or so wills are important for certain types of social historical studies (as witness the work on Nahuatl wills in "The Testaments of Culhuacan," "Beyond the Codices," and "Testaments of Toluca") I am not sure that the impact warrants the development of an online resource of parsed and glossed Zapotec wills marked up according to TEI (if that is the ultimate goal). Finally, there is the question as to whether the lexical resource created for interlinearizing the Doctrina through FLEx will contain the lexicographic data necessary to achieve the same result with the wills. Clearly it could be enhanced and the researchers on this project have the necessary knowledge of Zapotec to do so. Again, though, greater explanation of the lexical database for parsing through FLEx would have been helpful.

In sum, the team of researchers is eminently qualified and the material they have targeted, particularly the Doctrina, is extremely to the field of Zapotec linguistics. I think, however, that the proposal has not clearly expressed the mechanisms by which interlinearization will be carried out and contemplated whether the lexical resource itself might be a significant contribution.

**Overview**

Intellectual significance: For studies of Zapotec this is an important project.

Impact on research and technology: The technology is not new though the impact on research on Zapotec (historical and modern) would be significant.

Innovation: The creation of an online resource for colonial Mesoamerican texts, if done properly, could inspire similar work in the other major languages of the area: Nahuatl, Mixtec and Yucatec Maya.

Proposal development: These is insufficient attention to the nature of the lexical database for interlinearization through FLEx and insufficient justification for translation of the Arte into English.

Feasibility: FLEx can accomplish interlinearization though it is able to do so only if there is a lexical database with all the roots and morphemes that will be parsed out and glossed.

Qualifications: The team is highly qualified.