Ricky Reyes
Dr. Clovis Gladstone
Digital Texts
14 March 2024

Digital Texts Final Write Up (Part I)

Selection of Texts and Research Goals

For this project, I focused on digitizing and encoding collections of texts that document the history and cultural identity of Guam. While selecting texts, I immediately noticed a gap in the existing canon of digitized and archived materials about CHamoru history and experiences. As a CHamoru researcher and archivist, I have often encountered the challenge of accessing primary sources that center Pacifika perspectives. With this challenge in mind, I chose to reflect a range of narratives spanning texts authored by CHamoru authors to historical documents written by outsiders who have studied Guam through an academic and government lens. This distinction opened the doors for analysis of both internal community understandings around our traditions and history as well as an analysis of the discordance between how our culture exists and how our culture is understood by those who are outsiders.

In analyzing these texts, I witnessed first hand their important insights into historical events, political shifts, and cultural traditions that have shaped Guam over the last 300 plus years. Some of these texts are literary, incorporating poems, oral histories, and family stories, while others are academic or historical, offering analyses of colonialism, military occupation, and government accounts of historical land surveys. My goal in selecting these texts was to assemble a starting point for further research, ensuring that CHamoru perspectives are not lost within he broader discourse on Pasifika history and the cultural demolition currently happening under U.S. Military Occupation.

Digitization & TEI Encoding: Process & Challenges

Converting these texts into TEI required a more than automation of the markup process. Immediately, I noticed substantial challenges including the lack of a CHamoru large language model (LLM) and the diverse formatting of texts due to the diversity of textual data I'd chosen. For the first issue, unlike widely studied languages, CHamoru does not yet have the computational support necessary for natural language processing. This meant that I had to manually locate and encode CHamoru-language sections within the texts, making them using the TEI <lang> attribute. This approach ensured that future advancements in language processing could build upon this work, even as we deal with the current language restrictions we're working within.

For the issue of dealing with largely diverse formats of textual data, establishing a consistent TEI structure was proved to be a challenge early on due to the unique layout of each text. Since they varied in structure with some being literary works, other being non-fictions, and some poetry- the encoding process required significant time and flexibility. To address this issue, I implemented a structured TEI header for each work to ensure metadata uniformity (along with the addition of two unique tags <affiliations>, Insider/Outsider and <place>, Guam/Marianas/Other) and wrote a separate script to wrap the body contents in <p> tags as a base-level automated encoding measure. This approach still proved to be difficult for handling variations in front matter, back matter, and structural elements like footnotes which had to be, eventually, manually adjusted to ensure proper representation within the TEI file.

This process revealed the complexities of working with historical and literary texts that are genre-unique and did not conform to a single structural standard. Each document required different degrees of manual editing and no two texts followed the exact same encoding approach. In some cases, I adjusted TEI elements dynamically and on the fly to accommodate the text's format, ensuring the integrity of each work was maintained and, in other more simple cases, a simple script to wrap body text was sufficient.

PhiloLogic Analysis and Lessons

Using PhiloLogic to analyze this corpus gave slight insights into recurring themes, linguistic patters, and shifts in narrative perspectives across the shared texts. Unfortunately, due to the effort required to locate, digitize, and encode a diverse array of texts and due to the realization that the overall dataset of digitized CHamoru texts written by CHamoru people is fairly small, I didn't uncover as many broad patterns through PhiloLogic as I initially had hoped. This limitation itself is a significant finding. The scarcity of digitized Chamoru texts highlights ongoing issues of historical erasure and the urgent need for more preservation efforts if we want to be able to use distant reading tools like PhiloLogic to support any definitive declaration about Guam and CHamoru people based on this method of analysis. To improve search results, I applied lessons learned from part two if this project using TF-IDF, sentiment analysis, and topic modeling to create interesting search queries to guide my research questions and process. I am confident that with the addition of more texts and the utilization of these analysis tools, more interesting insights and questions will emerge. I am currently working with fellow CHamoru scholars to continue to add to this project and source texts - whether digitized or not.

Conclusion (Part I)

Concluding Part I, throughout this project I've taken what feels like an important step toward making historical and cultural texts about Guam more accessible for analysis within a structured digital framework. The challenges I faced, specifically the need for extensive manual encoding and the lack of CHamoru language models, highlight the complex nature of digitizing materials from underrepresented regions. However, this process has also demonstrated the value of TEI and PhiloLogic in preserving, standardizing, and analyzing these texts for future research within the current confines of our academic frameworks. By encoding these materials, my hope is that I can contribute to the ongoing efforts to document and share CHamoru history in ways that respect its complexity and lift the scholarship of CHamoru academics. This small corpus will, hopefully, serve as the start for further research enabling those interested to explore themes of colonialism, cultural resilience, and historical documentation in a structured and systematic way. As I continue to analyze the corpus and work with my fellow CHamoru friends, family, and scholars, I anticipate discovering new insights and intend to continue adding new texts to underscore the importance of encoding and analyzing these works to make Indigenous Pasifika histories more widely available in our digital archives.