Restoration of The General/Universal History of the Things of New Spain by Bernardino de Sahagún

Yash Hatekar

Indiana University Bloomington
yhatekar@iu.edu

December 12, 2022

Abstract

*The General/Universal History of the Things of New Spain comprises twelve books by the Franciscan Fray Bernardino de Sahagún (*1499 - †1590). After being lost for three centuries, a rediscovered manuscript of this ethnographic research study was named Florentine Codex after its storage in the Biblioteca Medicea Laurenziana in Florence. Composed of 12 books, it has the special and unique feature of being bilingual. Sahagún's manuscript was recognized as a World Heritage by UNESCO. And he is called the father of modern ethnography. Luckily Sahagún had given his promoter Fray Rodrigo de Sequeira a third copy (usually identified as the "Florentine Codex"), which Sequeira took to Spain in 1580. The Florentine Codex is the best-preserved manuscript of Sahagún's piece of work, kept in the Laurentian Library in Florence since 1783. In this project we are trying to restore this ancient text using natural language processing techniques.*

I. Introduction

Bernardino de Sahagún taught at the Colegio de Santa Cruz de Tlatelolco, the first European school on the American continent after the fall of the Aztec Empire (1519–21). [3] His students were young men from the Nahua nobility who helped him create a masterpiece of natural history based on interviews with indigenous informants in Tlatelolco, Texcoco, and Tenochtitlan. Unfortunately, in 1577, his works were confiscated by royal order. His research on the Aztec world was considered dangerous for the Conquest by giving knowledge to the indigenous people. Sadly, when Sahagún died, the Colegio de Santa Cruz deteriorated. By the 17th century, the rules imposed by the crown prohibited indigenous people from studying and teaching. There have been some improvements. But nowadays, there is still a lot to do to save the indigenous heritage. Since the Spanish conquerors destroyed almost all Aztec manuscripts dating before the Spanish Conquest, the work of Sahagún is the main and most often the sole source for studying and understanding central Mexico's civilizations before the Conquest. Composed of 12 books, it has the special and unique feature of being bilingual.

II. Dataset

The Florentine Codex is a complex document, morphed and appended over decades. Sahagún's goals of orienting fellow missionaries to Aztec culture, providing a rich Nahuatl vocabulary, and recording the indigenous cultural heritage are at times in competition within the work. [6] [7]Diverse voices, views, and opinions are expressed in these pages, and the result is a document that is sometimes contradictory. The Codex follows the organizational logic found in medieval encyclopedias,

in particular the 19-volume De proprietatibus rerum of Sahagún's fellow Franciscan Friar Bartholomew the Englishman. The Florentine Codex consists of twelve books. Each book in itself is different than the rest of the books, varying in lengths as well as the way of writing. The twelve books of the Florentine Codex are organized in the following way:

- Gods, religious beliefs and rituals, cosmology, and moral philosophy,
- Humanity (society, politics, economics, including anatomy and disease),
- Natural history.

The codex is composed of the following twelve books:

- The Gods. Deals with gods worshipped by the natives of this land, which is New Spain.
- The Ceremonies. Deals with holidays and sacrifices with which these natives honored their gods in times of infidelity.
- The Origin of the Gods. About the creation of the gods.
- The Soothsayers. About Indian judiciary astrology or omens and fortune-telling arts.
- The Omens. Deals with foretelling these natives made from birds, animals, and insects in order to foretell the future.
- Rhetoric and Moral Philosophy. About prayers to their gods, rhetoric, moral philosophy, and theology in the same context.
- The Sun, Moon and Stars, and the Binding of the Years. Deals with the sun, the moon, the stars, and the jubilee year.
- Kings and Lords. About kings and lords, and the way they held their elections and governed their reigns.
- The Merchants. About long-distance elite merchants, pochteca, who expanded trade, reconnoitered new areas to conquer, and agents-provocateurs.
- The People. About general history: it explains vices and virtues, spiritual as well as bodily, of all manner of persons.
- Earthly Things. About properties of animals, birds, fish, trees, herbs, flowers, metals, and stones, and about colors.
- The Conquest. About the conquest of New Spain from the Tenochtitlan-Tlatelolco point of view.

Examples of text from the Florentine Codex :

- de los dioses fo.1
  I nic ce amuxtli, vncan motene oa,in te teu.h: in qujnmoteutiaia in nican tlaca (1) – Book 1 fo.1
- c ei tenochtitlan tlatoani.
  Chimal popoca ic ei tlato cat in tenochtitlan matlac xiuitl.(5) – Book 8 fo.1

III. Preprocessing

To extract the page titles and page numbers from the books, we had to perform some preprocessing steps like :

- Adding 'de' to every title/page start if it was not present.
- Adding 'fo.xx' where xx is a number at the end of each title if it was missing to segregate the titles of the pages.
- Formatting the titles in the form 'de' 'abc' 'fo.xx', where abc is title of the page and xx represents a number.
- If a page didnot have a titles a new title in the form of 'de' 'fo.xx' was added.
- Adding a '.' in the beginning of every 'de' in title and in the end of every page number. This is done to simplify extraction of page title, page number and content of the page.

IV. Methodology

Restoration of this ancient text starts with extracting page wise sentences. We are using NLTK's [2] sentence tokenizer to extract page wise sentences in a .tsv file. The format in which each sentence is stored is 'page_title' 'fo.' 'number' 'sentence'. We observed that several words have spelling mistakes/fallacies and need corrections. For example - [u] -> o , [yj] -> i and [iao] -> yao . So we have used a python library Py-Elotl [1] to generate

the normalized forms of these incorrect words. The python library Py-Elotl provides us with three different normalizations :

- sep
  - Alphabet often seen in use by the Secretaría de Educación Pública (SEP) and the Instituto Nacional para la Educación de los Adultos (INEA). important characteristics of this alphabet are the use of "u" for the phoneme /w/, "k" for /k/, and "j" for /h/.
- inali
  - Alphabet in use by the Instituto Nacional de Lenguas Indígenas. Uses "w" for /w/, "k" for /k/, and "h" for /h/.
- ack
  - Alphabet initially used by Richard Andrews and subsequently by a number of other Nahuatl scholars. Named after Andrews, Campbell, and Karttunen. Uses "hu" for /w/, "c" and "qu" for /k/, and "h" for /h/.

The library uses sep noramlization by default. We have also introduced some custom rules to normalize the incorrect words. Following which we have added a new column with the normalized form of every sentence. The python script is run for all the twelve books individually after each book has undergone the required preprocessing.

V. Discussion

Example of the output of the python scripts :

——— ———

.de Quavitleoa. fo.16. iotiaia quetzalxoch, y ni tlatquj catca texotic. J nic nauhcan, poiauhtla, canitzintla, çan ixpan in tepetl, te petzinco: itoca ietiujia in miquja poiauhtecatl, y nic muchichiuh tiuja, vlpiiaoac, tlaulujtectli. J nic ma cujlcan, vmpa in atl itic itocaiocan pantitlan, in vmpa, on mj quja, itoca ietiuh, epcoatl: y ni tlatquj, in ca qujtiuja, epne panjuhquj. J nic chi qua ceccan: vm pa qujujcaia cocotl icpac, no itoca ietiuja cocotl: y ni ne chi chioal cat ca, chic tla pan quj, cectlapal chi chiltic, cectlapal y iappalli. J nic chicoccan, icpac y nj iauh queme: can no itoca ietiuh, y iauh queme, in tlacateteujtl: y nj tlat quj ieti uh, tlacemaqujlli, y nj ia ppalli. Jz quj çan yn, in mi qujia nextla oalti, tlacateteuhti: auh mu chinti, yn maxtlatzon ietiuh, que tzalxixiquj, quetzal mj iaoaio: in chalchiuhcozquj ietiuh, yoan mo ma cueztitiuj, quj moma cuextiti uj chalchiujtl; tlaixolujlti, quj mixolhujaia, mjx mi chioaujque, yoan y molcac, y molcac ietiuh, muchintin mauizcotiuj, tlacenca oalti, tlachichioalti, muchi tlaco tlanquj, y njn tech ietiuh, tlaco tlantiuj, qujmamatlapaltia, amatl, ama amatlapaleque: tlapechtica in vicoia, quetzalcallotiuja, yn vnca momantiujia, qujntlapichilitiuja: cenca tlatlao culltiaia, techoctiaia, techoquizvitomaia, teicnotlama chtiaia, ynca elciciooaia. A uh yn oaxitiloque tococan, aiauhcalco, vncan ce iooal tocaujlo, qujntoca viia tlamacazque: yoan in qua quacujlti, iehoan in ie veuetque tlamacazque. A uh in tlamacaz que, y noceme ontetlalcaujque, motocaiotiaia mocauhque: aoc mo tecujcananamiquj, aoccan onmonequj, aoccan onpoalo.(2) A uh in pipiltzitzinti, intla choca tiuj, intla imixaio toto catiuh, in tla imixaio pipilcatiuh, mjtoaia, moteneoaia, ca quj iauiz: y ni mixa io qujnezcaiotiaia, in qujiaujtl, ic papacoaia, ic teiollo motlalia ia: iuh qujtoa, ca ie moquetzaz in quj iaujtl, ca ie tiqujiaujlozque. Auh intla cana ca, y tixiuhquj, q toaia, amo techqujiaujlotla. A uh yn ie qujcaz qujiaujtl, yn ie tlamiz y nie itzonco: njman ie ic tlatoa in cujtlacochi, y nezca, y nieujtz, in ic moquetzaz tlapaqujiaujtl: njman oalhuj, pipixcame: no yoa

——— Using elotl sep——— .de Quavitleoa. fo.16. iotiaia quetzalxoch, y ni tlatquj catca texotic. iotiaia ketsalxoch, i ni tlatquj katka texotik.

——— Using elotl ina——— .de Quavitleoa. fo.16. iotiaia quetzalxoch, y ni tlatquj catca texotic. iotiaia ketsalxoch, i ni tlatquj katka texotik.

——— Using elotl ack——— .de Quavitleoa. fo.16. iotiaia quetzalxoch, y ni tlatquj catca texotic. iotiaia quetzalxoch, i ni tlatquj catca texotic.

——— Using custom rules —- .de Quavitleoa. fo.16. iotiaia quetzalxoch, y ni tlatquj catca texotic. iotiaia qoetzalxoch, y ni tlatqoj catca texotic.

In the output displayed above we can see the normalized form for one of the sentence from the page taken from Book 2. Similar to this all the other sentence and its normalized form are printed below the original page from the book in a .tsv file.

From the outputs of our script we could observe that, most of the sentences are tokenized correctly. We could also see mostly correct normalized forms for each of the sentence. We could also see that if the words are split between lines the words are not captured properly, and hence are not normalized correctly. We could also see some sentences not being correctly tokenized.

VI. Conclusion

Restoration of such ancient texts not only tells us more about happenings during those ancient times but these texts become foundations of the religious beliefs, humanity and natural history. With better and upcoming natural language processing techniques it is possible to restore them so that common mass can also read and learn from them irrespective of the language these texts were written in.

VII. Future Work

Firstly we can improve sentence tokenization by using POS tagging and using Nahuatl sentence form grammar to accurately find the correct sentence boundaries. A method to club the words split between sentences needs to be added.

References

[1] Py-Elotl : https://github.com/ElotlMX/py-elotl

[2] Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit : https://github.com/nltk/nltk

[3] Florentine Codex : https://zenodo.org/record/7213020#.Y0wJl0zP1PY

[4] Bernardino de Sahagún. General History of the Things of New Spain by Fray Bernardino de Sahagún: The Florentine Codex. Book I: The Gods. [Place of Publication Not Identified: Publisher Not Identified, 1577] Pdf. Retrieved from the Library of Congress.

[5] Bernardino de Sahagún. Florentine Codex: General History of the Things of New Spain: Book 1 The Gods. Translated by Charles E. Dibble and Arthur J. O. Anderson. 2. Salt Lake City: The School of American Research and the University of Utah, 2012.

[6] Wikipedia : https://en.wikipedia.org/wiki/Florentine_Codex#Evolution,_format,_and_structure

[7] Youtube : https://youtu.be/FnQwfvP-QJE