TNA ACTIVITY REPORT

Distant reading of early modern prophecies across Europe: corpus formation and testing

Author: Eduardo Fernández Guerrero

Current position: European University Institute

Affiliation: Early Stage Researcher

Host institution: UNED

Mentor(s): Salvador Ros

Period of stay: 4.10.2023-20.12.2023

Introduction

In this project I attempted to set forth the foundations for a corpus of early modern prophecies in Latin written between Spain and Italy during the 16th century. Prophecies and revelations are a central – yet often overlooked – element of Western literature, and one that crosses linguistic, chronological, and religious divides alike. From political prophecies to visionary accounts, such texts sit awkwardly between literature and history: themselves part of literary tradition that imposes styles and topics inherent to fiction, they nevertheless claim authority and veracity while articulating alternative futures and challenges to the intellectual, political or religious order. Disregarded often as merely political propaganda or part of what medievalists have called the *pastorale de la peur*, only recently prophecies have started receiving attention on their own merits, not just as "means to an end" but rather as a meaningful cultural object of a given time. Moreover, prophecies and revelations are a particularly useful site to challenge traditional periodizations, as their continuity well into the modern period is starting to gain scholarly attention.

CLS INFRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984

Methodological plan

The project intended to create the framework for FAIR corpus of early modern prophecies in Latin. This implied a series of steps, from establishing the criteria for inclusion in the corpus, to developing a pipeline to automatically identify and retrieve relevant texts from existing corpora as well as from scanned books and manuscripts. Such texts were then to be orthographically standardized via developing a set of regular expressions. Finally, sentiment analysis and part-of-speech (PoS) tagging needed to be applied to understand emotional tones, linguistic structures, and thematic patterns within the prophecies.

Description of research visit and its outcome

The implementation of the plan above presented a series of challenges that I managed to solve and learn from thanks to the mentors at my host institution.

First, the selection of texts raised several issues: in order to create a homogenous corpus that allowed internal comparisons I established a chronology of 1400 – 1600. I had to rule out subsets such as astrological prophecies that would involve a different lexical fields, as these would not relate adequately to texts describing visions or extasies (the main focus of the project). The existing online corpora of Latin texts that were meant to be the starting point for the extraction of texts into my own proved also challenging for this reason, with most texts associated with terms such as ‘prophetia’ or ‘prophetare’ turning out to be either astrological or outside the chronological range established.

During my visit my mentors explained much about the concepts of bibliographic ontologies and relational databases, at the heart of the project. With this basis, and given the failure to automatize the input of texts into my corpus, I set out to do a review of the literature on early modern prophecies in order to identify texts matching the characteristics.

This led to a tentative list of titles that then I needed to retrieve and process in order to incorporate them into the corpus. Once downloaded from a series of online repositories (e.g. Google Books, BEIC, Internet Archive, etc), or scanned from printed editions, each individual book required processing. As the scanning settings for each book varies, a great deal of attention was paid into learning about the pipeline pdf > jpg > processed jpg > ocr > text, from the automatized download to the processing of jpgs (binarization with dynamic threshold, noise reduction, etc), in order to set up the basis for scaling up the process.

While I succeeded at developing a robust understanding of these processes, the degree to which they required human verification (selection of relevant pages within each book; identification of noise elements such as running titles, etc) made this a time-consuming task. This, together with the issues associated with the identification of suitable texts, took a significant part of my time. For each of the text, an entry in a spreadsheet was made, indicating a series of values for the text, aiming at both facilitating the identification of the original they were taken from, as well as providing key information about each text, such as provenance, themes, title, etc. Once I achieved a core of ca. 90 texts to work as proof of concept, I proceeded to the next steps, ie. morphological and PoS tagging with LatinCy. For this, it was crucial the support of my hosts, who developed with me the scripts to generate and visualize the results.

The outcome was satisfactory, although challenging. Two main questions animated the project in its inception: first, how the socio-political environment of the time influenced the content and style of prophecies; secondly, which recurring linguistic patterns or rhetorical devices could be identified in these prophecies that sets them apart of other genres. If the answer to the first could only be answered by a much larger sample of texts, answering the second one became quite revealing: not only there are some features to be expected, such as the prevalence of future tenses, but also some characteristics came through, such as the prevalence of Hebrew structures (e.g. using past tenses for expressing future times) in several strains of prophetic literature inspired in the Old Testament.

Considerations over future work

My project was originally inspired by the distant reading of contemporary literature, often involving readily available in great number. The application of such techniques and questions to an undefined corpus of early modern literature, for which often there are no modern editions presents some challenges. While the questions remain productive, and the outcome suggests the applicability of this framework, its true potential seems to be only accessible via a long-term, collaborative project.