
GLOBAL MEDIEVAL SOURCEBOOK
TEXT TECH
IMMIGRATION DETENTION
THE LITERARY LAB
JOSQUIN RESEARCH PROJECT
SAC MUSEUM CULTURES
SENEGAL SLAVE UBE RATIONS PROJECT
PAN PANDEMIC & C
MAPPING TTOMAN EPIRUS
CESTA
anthology
2020-2021
MEDIEVAL networks of EMORY
MODERNIST ARCHIVES PUBLISHING PROJECT
RUMSEY MAP CENTER CHEROKEE HISTORY
CISTERN
grand tour project
COMPUTATIONAL BORDER STUDIES
SOCIAL NETWORKS IN EARLY ISLAMIC MIDDLE EAST
GEORGE MOSES HORTON PROJECT
OPEN GULF
ORAL HISTORY TEXT ANALYSIS PROJECT
AFTERMATHS OF ENSLAVEMENT
LIFE IN QUARANTINE
URBAN STUDIES AND PRACTICE OF THEORY
POETIC MEDIA LAB
SOCIAL NETWORKS ROMAN COMEDY
GLOBAL URBANIZATION AND ITS DISCONTENTS

director's WELCOME

We at CESTA take special pride in our Undergraduate Research Internship program, through which humanities faculty and senior researchers work on collaborative project teams with students from all across campus. We are grateful that, despite the pandemic, these transgenerational collaborations have continued to thrive—as reflected in this year's CESTA Anthology, which describes our research projects from the point of view of our 2021 undergraduate students.

Our undergraduate researchers faced unique challenges during this time of uncertainty. What was to be a Virtual Summer of remote collaboration became a Virtual Year, and yet our students adapted and produced extraordinary work. In-person meetings this Autumn have been immensely touching, bringing together project members who had previously only worked together virtually. Although we were fortunate to be able to advance exciting research agendas on-screen, we missed the materiality of our human connections. The design of this year's Anthology, centered on doodles, speaks to this craving for the in-person and the handmade—for engagement that takes place off-screen.

It's hard to convey just how much effort, energy and creativity has gone into CESTA's work this year. We are grateful for our resilient community of students and faculty, and also to the tremendous staff that sustained us in 2021: Amanda Wilson-Bergado, Daniel Bush and Erik Steiner, who were joined for the Internship program by graduate mentors. A team of graduates and postdocs worked tirelessly on this Anthology through the Autumn. More than ever we wish to thank the Dean of Research, the Vice Provost for Undergraduate Education, the H&S Dean's Office, the Stanford Humanities Center, and our generous donors, all of whose support is critical to everything that we do. Please enjoy the pages that follow, which tell the story of what we do here at CESTA in our brilliant students' words.

CESTA FACULTY DIRECTOR
Giovanna Ceserani

The 2020-21 academic year was truly unprecedented. We started online courses in March 2020 and continued our scholarship, teaching, and research virtually from locations across the globe. From Stanford, Virginia, Arizona, New Jersey, and Turkey, the graduate mentor team collaborated across time zones and digital platforms to help bright CESTA interns transform ideas into reality.

note from the MENTOR TEAM

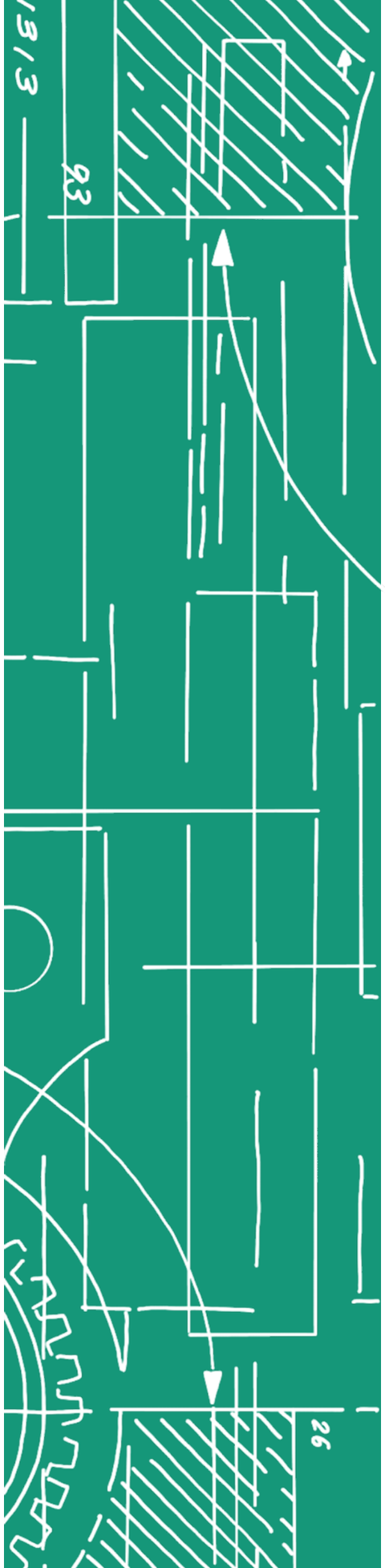
Merve: When I joined the CESTA team as a graduate mentor in Spring Quarter 2021, I was not sure what to expect. I knew only a handful of the interns and had only a rough idea about the research projects. During spring and summer, however, I got to meet some amazing students with bright ideas, the energy to turn them into reality, and the compassion to work with and alongside colleagues despite the added challenges of the long pandemic. Meeting with interns regularly, hearing about their progress and working through their questions was not only a job that I thoroughly enjoyed but also one of the highlights of my week. I welcomed the student interactions amidst the habitual loneliness of my research and cherished seeing projects grow in new directions.

Annie: I had the honor of working as CESTA's graduate technical mentor over the summer quarter. The summer was unique for so many reasons—our students were scattered across the globe and research materials were limited. In spite of these challenges, our interns demonstrated quality research skills and were eager to learn more than what was required for their projects. Several requested programming tutorials, introductions to data science, and advice about database security. Working with interns, faculty leads and researchers, and watching their projects develop, was the highlight of my summer.

This past academic year was arduous for all of us, and the fact that we did so much together at CESTA is a testament to the strength of our community. This Anthology is a record of the past year and a symbol of our resilience.

GRADUATE STUDENT MENTORS
Merve Tekgürler and Annie Lamar

TABLE OF CONTENTS



Stanford sits on the ancestral land of the Muwekma Ohlone Tribe. This land was and continues to be of great importance to the Ohlone people. Consistent with our values of community and inclusion, we have a responsibility to acknowledge, honor, and make visible the University's relationship to Native people.

OPENGULF:
HISTORICAL TEXTS AS DATA

Project Description

Nora Barakat,
Assistant Professor
of History

OpenGulf (opengulf.github.io) is a transdisciplinary, multi-institutional research group analyzing historical texts produced in the Arabian Peninsula, Iran and Iraq from the early nineteenth century to the present. The various projects associated with OpenGulf publish open historical datasets, corpora and digital exhibitions with the aim of opening the field of Gulf Studies to digital historical exploration, analysis and interpretation in the service of open research and pedagogy. Currently, OpenGulf includes six projects with students, faculty and staff at eight institutions actively contributing content including handwritten text analysis of Arabic, English and Ottoman Turkish texts, interdisciplinary analysis of data from phone books from the city of Abu Dhabi from the 1970s-2000s, and close and distant readings of an expansive British gazetteer of the region that includes mapping over 20,000 unique named locations. During the 2020-2021 academic year, CESTA interns worked on the Historical Texts as Data project of OpenGulf, which follows a general three-step workflow: preparing historical texts in various media formats and languages for digital analysis; extracting and annotating names of people and places in those texts to create reusable structured data; and creating and publishing visualizations and narratives derived from those datasets.

Better Training Sets, Better Research

Moe Khalil

My work with OpenGulf has consisted of several distinct yet interrelated projects. Most of my research has used Transkribus, a platform that provides language-independent, AI-powered handwritten text recognition to translate texts from digitized image format to searchable, computationally-tractable text. Unfortunately, due to the limited Arabic training sets in Transkribus, the existing Arabic text recognition was often incorrect. Furthermore, all available Arabic models had been trained on old newspapers or printed editions, which usually do not include the diacritics and punctuation found in larger literary texts. These training sets were useful as base models, but we needed to include new domains of texts in order to retrain the model.

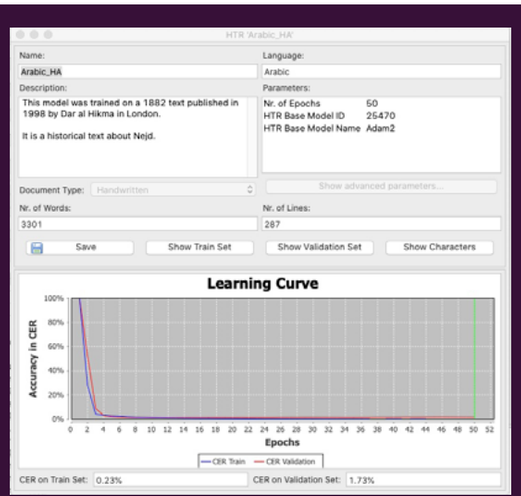


Figure 1: This screenshot from Transkribus shows the change in accuracy in Arabic text recognition following our expansion of the training dataset. The increased prevalence of literary texts allowed us to increase the accuracy (CER Validation) in relation to the test data (CER Train), thereby making the model operate better with materials from Lorimer.

Initially, we focused on printed texts that were not available in text form. For example, we first used an 1882 gazetteer of Baghdad, Basra, and Najd written by Al-Haydari. This text included numerous place names and geographical descriptions that would be valuable for the larger scope of the project. Throughout much of my time with OpenGulf I worked on training a model through Transkribus in order to be able to recognize and transcribe such text, and was eventually able to bring the error rate of the model down to around 1.7%—a great improvement.

After this success, we decided to train the model to recognize purely handwritten texts. To accomplish this, we transcribed several handwritten texts from various sources, including a set of accounting letters and a travel journal. In the future, we plan to create a more comprehensive data set that can be used as the starting point for an Arabic handwritten text recognition model in Transkribus.

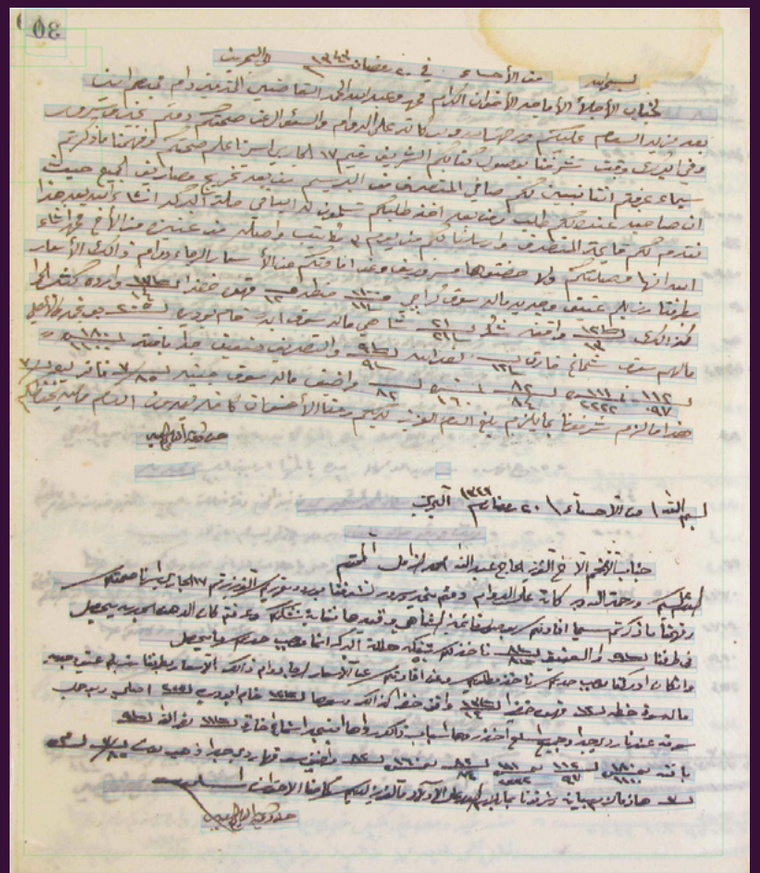


Figure 2: This handwritten Arabic letter from April 1925 is one of many sourced from the collection of the Bahraini merchant Muhammad al-Ajjaji, held in the Juma bin Majid library in Dubai, United Arab Emirates. The light blue text boxes visible in the image are automatically generated through Transkribus's layout analysis algorithm.

Computing Lorimer

Danny Tse

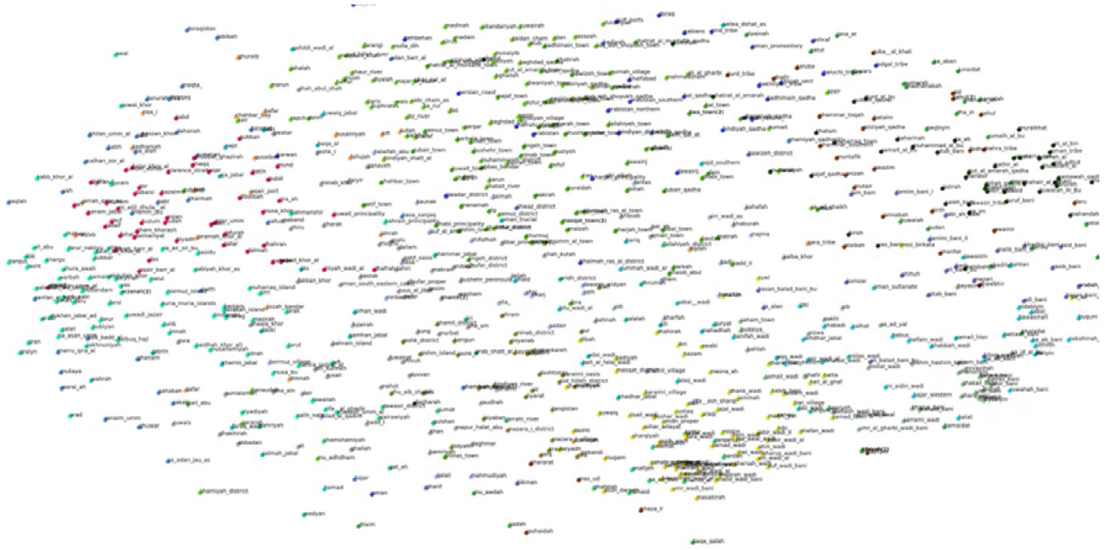
As part of my research for OpenGulf: Historical Texts as Data Project, I consulted with field experts Prof. Nora Barakat and Prof. David Wrisley to compile questions related to the Gazetteer of the Persian Gulf, Oman, and Central Arabia (henceforth, Lorimer) that could focus my research. Ultimately, the primary questions guiding my research centered on whether or not words in a corpus can be mapped onto regions as we know them, or if they create clusters that are other than geographical, and whether we can analyze them to predict the common themes of the chapters they're found in.

GEOGRAPHICAL AND STATISTICAL GAZETTEER
OF THE
PERSIAN GULF,
'OMĀN,
AND CENTRAL ARABIA.

Singular 'Abdūli عبدولي : a tribe of Trucial **'Omān** who have 200 houses at **Shārjah** Town, 20 at **Ghāllah** in **Shamailīyah**, and 15 at **Khalaibīyah** adjoining Wādi **Hām**; some of them are found also as settlers on Shaikh **Shu'aib** island. In all they may number about 1,200 souls. In politics they are Ghāfiris, and in religion Hanbali Sunnis. They are not connected with the **Sharqiyīn**, and they claim, it is said, to be Shurafa from Makkah. Another account assimilates them to the 'Obaidli tribe of the **Shībkuh** district in Persia, deriving them from the 'Abdah branch of the **Shammar** of **Najd**; it is possible that this theory has no foundation except in the partial resemblance of the names. **'ABĀDILA** عبادلة

Figure 3: Example of a page from a chapter in Lorimer that serves as our research material.

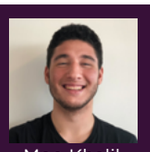
Initially, it was difficult to read through each text and categorizeit based on predicted geographic location. As a solution, I first began by writing a Python script to generate a spreadsheet of top-n words appearing within each text. Then, I analyzed the results and removed common stop words such as articles and basic adjectives. In our group, we pondered whether to remove distance words, such as "miles" and "feet." On the one hand, these could help us estimate the size of the region being described. On the other, these were largely function words that blocked more important and meaning-bearing words from being displayed. Though the initial attempt gave us a good amount of information to work with, we realized that we still had to manually cluster chapters based on their top words. Thus, I adapted and deployed a tf-idf-based model on the text, hoping to visualize the distribution of texts in a more legible manner.



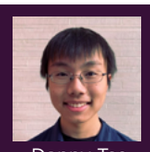
Term frequency-inverse document frequency, or tf-idf, is a score that showcases the significance of a word within its document in a corpus. Essentially, if a word appears more in a specific document, then the tf-idf score increases. Conversely, if a word appears just as frequently throughout the corpus, then the td-idf score decreases. In my model, I used a form of tf-idf imported from sklearn. feature_extraction.text, which generates a matrix from counting word occurrences per document. Then, I apply weights within this matrix as described above: words that appeared frequently within a document but not throughout the corpus received higher scores.

I then utilized *k*-means clustering, in which we cluster our texts into *n* clusters. In short, the algorithm assigns each text to a cluster with the closest centroid, then recomputes the cluster centroid as the mean of all the texts in the cluster. Since there were over 800 documents, I tested both $n = 10$ and $n = 20$, and using multidimensional scaling, mapped the texts into a 2D grid where texts were represented as points color-coded based on their respective clusters. However, we found that this visualization was difficult to read, so I explored other clustering/visualization algorithms, such as hierarchical document clustering. Future work will include developing visualizations that are easier to read and experimentation with other types of models.

STUDENT RESEARCHERS



Moe Khalil



Danny Tse