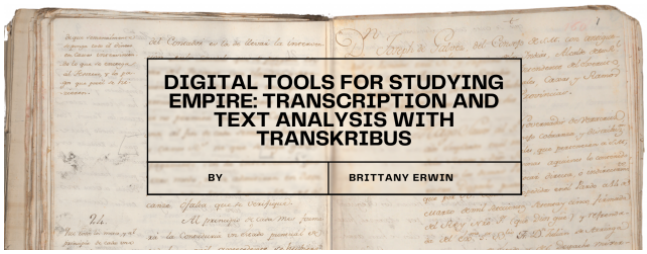
Digital Tools for Studying Empire: Transcription and Text Analysis with Transkribus



By Brittany Erwin

Transcribing historical documents, that is, copying the exact text word-for-word, can be a long and arduous process. Some handwriting styles from earlier periods have become almost indecipherable to modern-day readers without extensive training and practice. Many texts also feature abbreviations, which helped keep the cost of paper and ink down for the original authors. The meaning of those shortened words and phrases may have been commonplace centuries ago, but they are far from self-evident today.

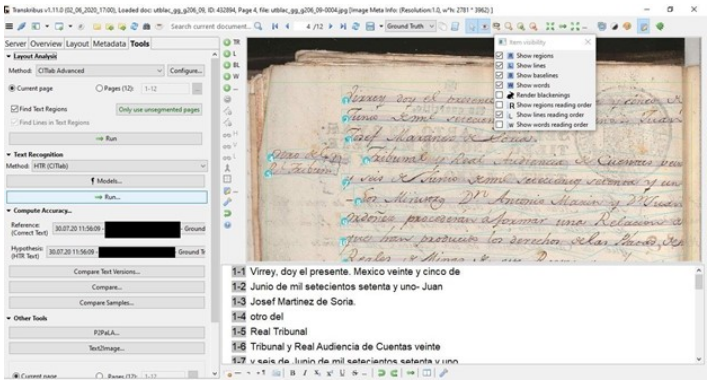
Although it requires considerable effort and difficulty, transcription is often essential for historical research. Digital versions of historical texts allow for analysis on a larger scale because they transform each document into a searchable field. For example, identifying trends, common words, and shared phrases across dozens of documents enables a significantly quicker process.

Within the realm of Digital Humanities, many tools exist to help facilitate transcription, with even more currently being developed. Transkribus is one useful interface. I utilized the tools that Transkribus offers to analyze the frequency of prominent phrases in around thirty documents from the Genaro García Collection, housed at the Benson Latin American Collection at UT-Austin.

The Genaro García documents I transcribed with Transkribus originate from the interactive digital exhibition that I created on the 1765 *visita*, or royal inspection, of New Spain. The *visita* examined local institutions, evaluated economic policies, and reorganized society in a broad display of royal authority. This procedure helped the reigning monarch (Charles III) implement widespread political, economic, and social reform in this territory in order to tighten control and increase efficiency. It set the precedent for changing policies throughout the empire over the next several decades.

Designed for a non-specialist audience, the exhibition explores the timeline, spatial breadth, and procedure of the inspection, by providing access to digital versions of the original documents produced by the royal inspection *visita*. The project provides an accessible way to understand how the lengthy and expensive process of royal governance effectively fostered relations between the ruling government in Spain and its many different constituencies on the ground in the Americas. I prioritized transcription of the *visita* documents to help shed light on the Crown's objectives for imperial reform.

Transkribus is a fairly new platform, designed by the Digitisation and Digital Preservation Group at the University of Innsbruck in Austria. Its basic function is creating programs that learn to read documents and produce transcriptions on their own. The user interface allows researchers to design the programs, monitor their progress, and correct them as needed.

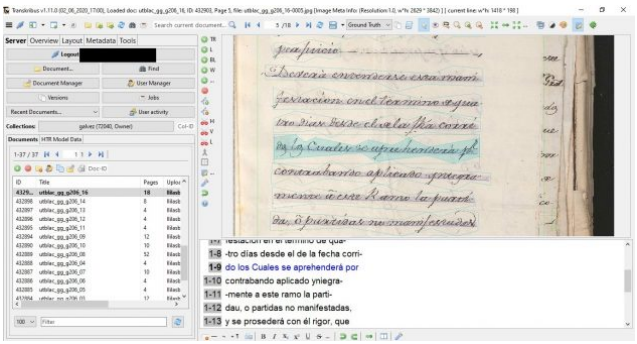


The Layout Analysis feature appears at the top of the panel on the right side of the screen. Users can run Layout Analysis one page at a time, or one document at a time.

The first step in the process is to upload high-resolution images of the documents in question and employ the Layout Analysis feature. This step automatically maps out the lines of text on each page so that the next feature, the actual transcription, knows where to look for the characters. It will only read letters within those lines. The user can manually add or delete lines in case the Layout Analysis feature made a mistake, such as recognizing a spare mark or an ink blot as text.

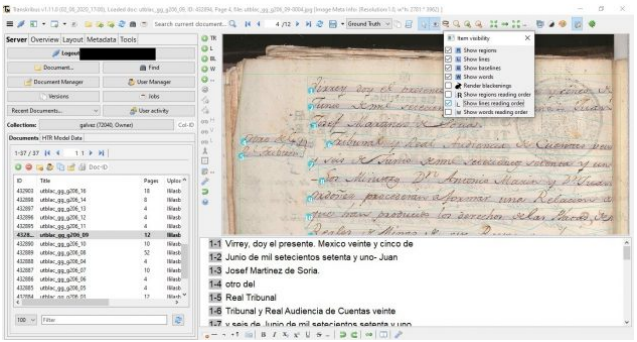
After Layout Analysis is complete, users can begin to build their transcription model. Transkribus models work based on an input from the user. In most cases, a manually transcribed document of about 15,000 words serves as the ideal input. The program will essentially learn to match characters in the image of the document to the ones provided in the manually-created transcription. From there, the model can read and translate any number of subsequent documents. The wait time for transcription depends on the size of the document, but it can be ready in as soon as a few hours.

An important point about the efficacy of these Transkribus models: handwriting matters. Since the program will read documents based on the model, it is essential that the handwriting for both is either the same or highly similar. For my specific documents, the handwriting was consistent, coming mostly from a Spanish Royal Inspector named José de Gálvez.



The above image illustrates that the highlighted line of text matches the line below, labeled 1-9, where the user can edit the text as necessary.

Even after the program generates transcriptions for the selected documents, the Transkribus interface allows for a manual review process. For example, if the model misinterpreted the flow of the text, users can reorder the lines. They can also correct any misread words by simply clicking on the relevant line and replacing the letters or characters. At any point in this process, users can save their work, which Transkribus automatically backs up on its own servers.



The box in the top right has the option, which is currently selected, to "Show lines reading order. If the user wishes to reorder the lines, she can double click on any of the light blue numbers to change it.

There are a few options for the end product of a document transcribed with Transkribus. Users can download integrated files, which include the photo of the document together with its text transcription, a text-only file (.txt), or a Word document (.doc). Each of those formats provides a digital version of the original document that is well-suited for continued analysis.

One platform that works well with a .txt file is Voyant Tools. By simply uploading the text (or even copying and pasting it) in Voyant, users can track the frequency of certain words and create data visualizations, such as word clouds and graphs. Especially for a large collection of long documents, such visual representations can lead to new, unexpected insights.

This word cloud, called a Cirrus in Voyant, was one of the data visualizations that I created from my transcriptions from Transkribus.

Transkribus is a promising tool for researchers working in many different languages and time periods. Its ability to create a model based specifically on the handwriting of the documents that a user selects make it a highly adaptable platform. It is currently offered as a free interface, but is transitioning to paid service that would charge based on the file-size of projects. However, the developers have prioritized accessibility, and aim to offer scholarships to undergraduate and graduate students.

Posted November 6, 2020

More Digital & Film

We are supported by the College of Liberal Arts
And our **READERS**

DONATE

CONTACT

Sign up to receive the monthly Not Even Past newsletter

Your email address

SUBSCRIBE