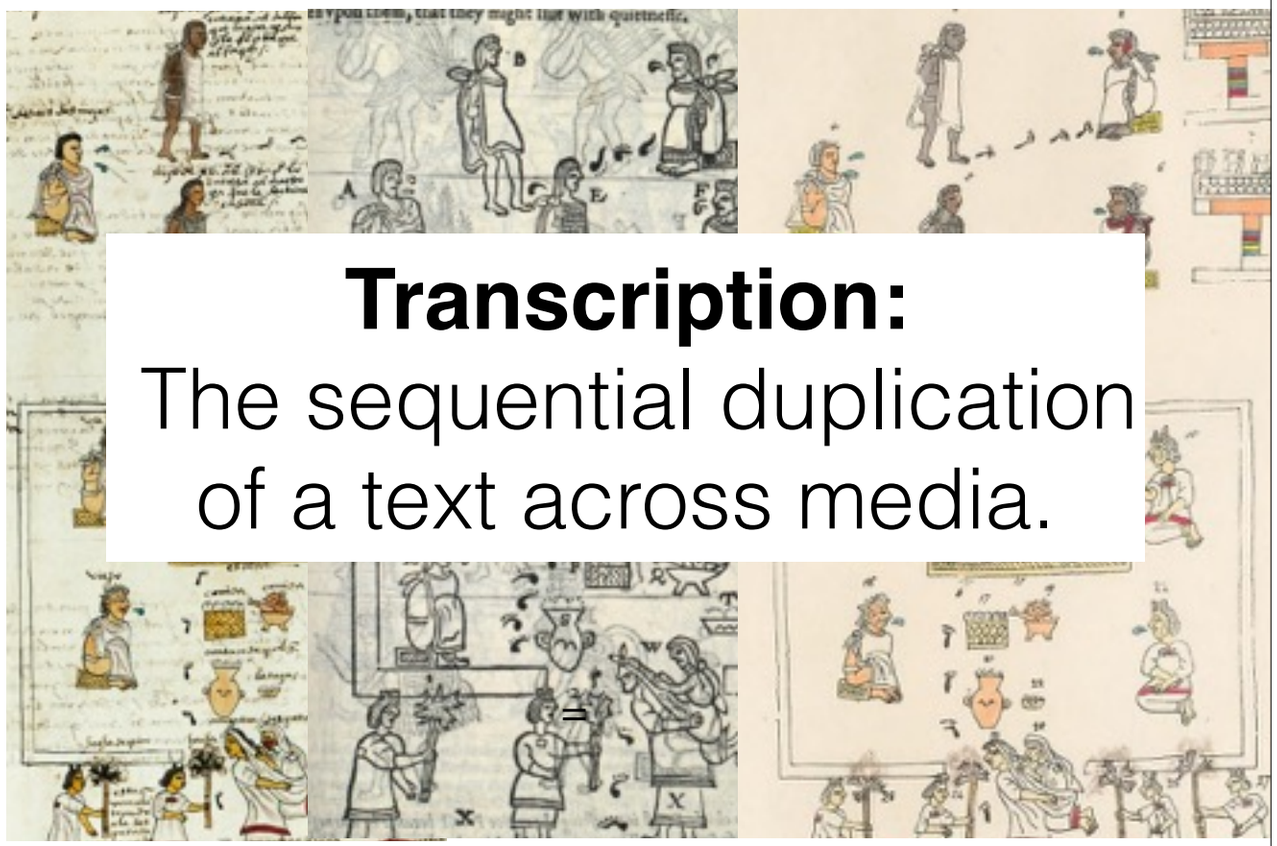
Esta ãphibologia no ay ẽ latin

Machine reading linguistic hybridity in the Primeros Libros

Hannah Alpert-Abrams
University of Texas at Austin

Good morning. My name is Hannah Alpert-Abrams and I am a PhD candidate at the University of Texas at Austin. My talk today is about the transcription of historical documents.

Definition



To begin, I'd like invite everyone here to think back to the last time you transcribed something. Maybe you conducted interviews recently, and have been transcribing them into typed text. Maybe you were visiting the archivo general and copying manuscripts. Or maybe you were copying text from a slide onto your laptop. Or typing a quote from a talk into twitter.

Scholars transcribe. Scholars transcribe all the time, in all parts of our lives, but we rarely give sustained attention to the act of transcription: the sequential duplication of a text across media. Like markup or typesetting, transcription is one of the many small labors of scholarly practice that go largely - though not completely - unnoted. With this talk, I seek to turn critical attention to this practice.

Reading the First Books

Automatic transcription for multilingual, early modern printed documents

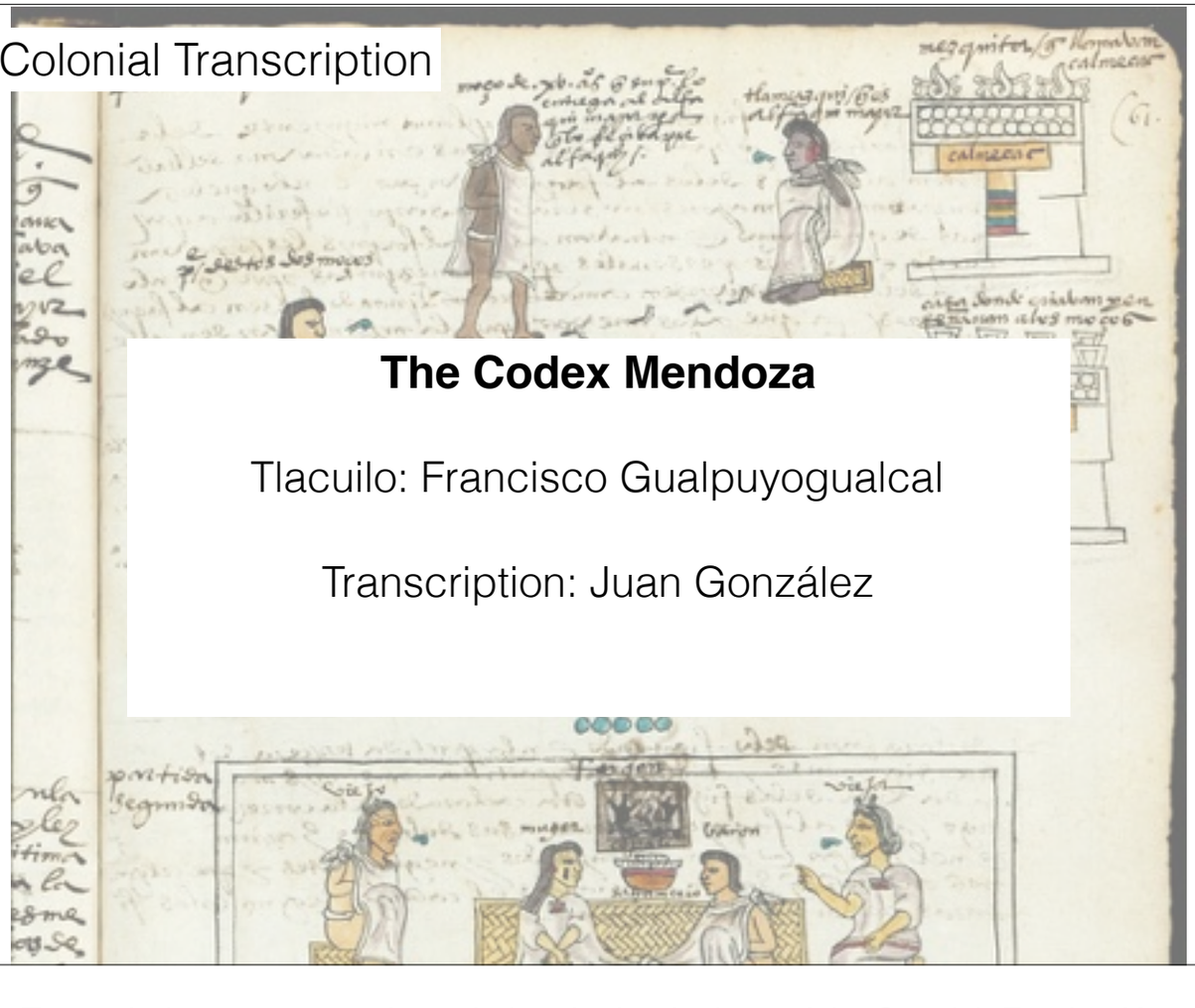
sites.utexas.edu/firstbooks/

The point of departure for this talk is the Reading the First Books project, which is an effort to produce tools for the automatic transcription of multilingual, early modern printed books. In the second half of this talk I'm going to talk about the project, which is beginning its second year. I hope that those of you who do scholarly editing or digitization projects will find this informative.

Before I do that, however, I want to take some time to think seriously about what it means to transcribe automatically. The Reading the First Books project is focused on the challenges to transcription posed by digital facsimiles in the Primeros Libros collection of books printed before 1601 in the Americas. Specifically, we are concerned with language use and orthography.

This is not the first time that Early Colonial American documents have been transcribed, however, or that intellectuals of various kinds of sought to handle challenges of language and orthography. Before I talk about the project, then, I want to situate it within a longer history of transcription that begins in the sixteenth century and extends to the present day. This is not to say that there is a linear story, and even if there were, I wouldn't have time to tell it. Instead, I'll just offer a couple of brief vignettes that I hope illustrate some of the challenges of working with transcribed texts.

Colonial Transcription



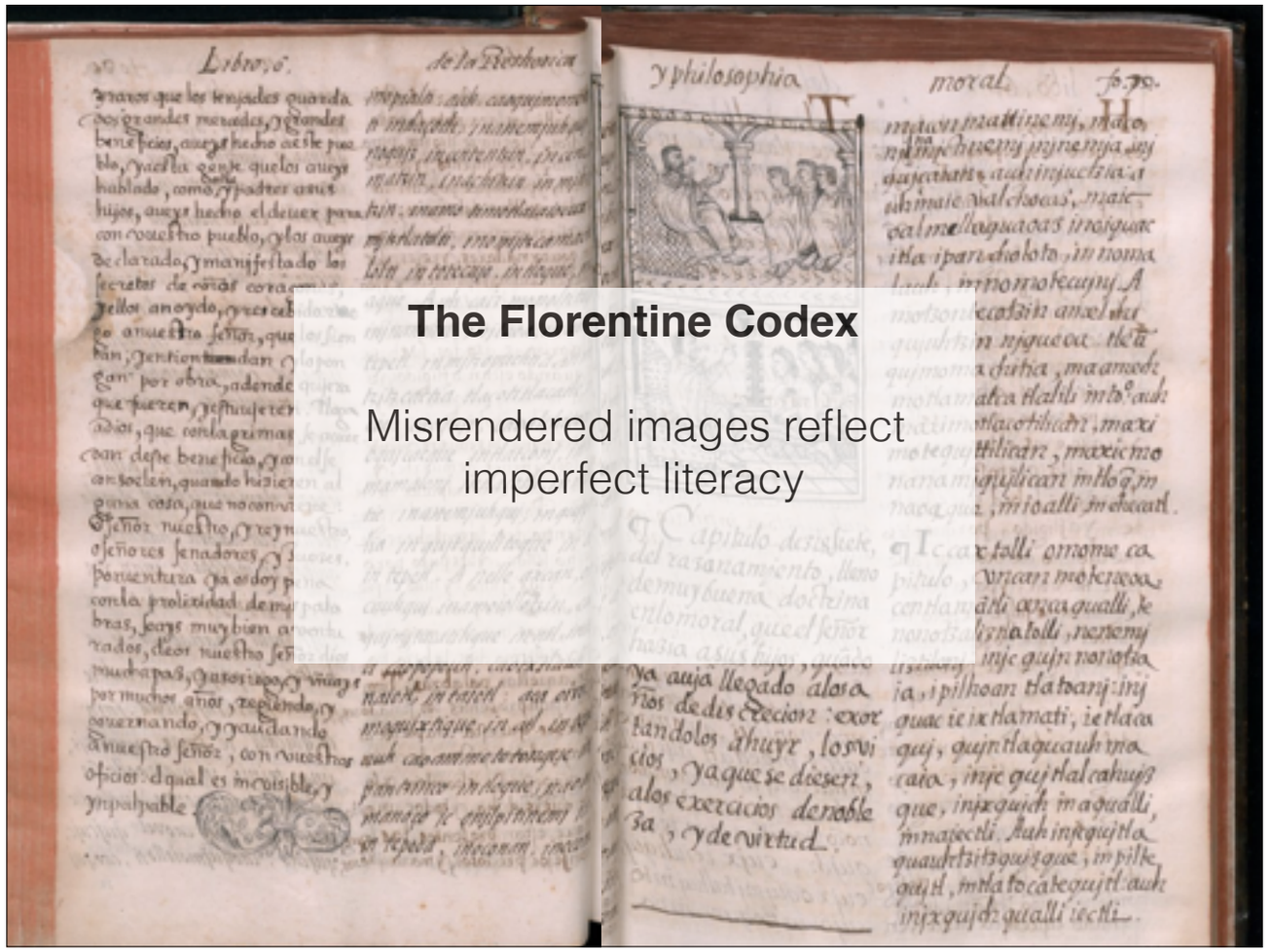
The Codex Mendoza

Tlacuilo: Francisco Gualpuyogualcal

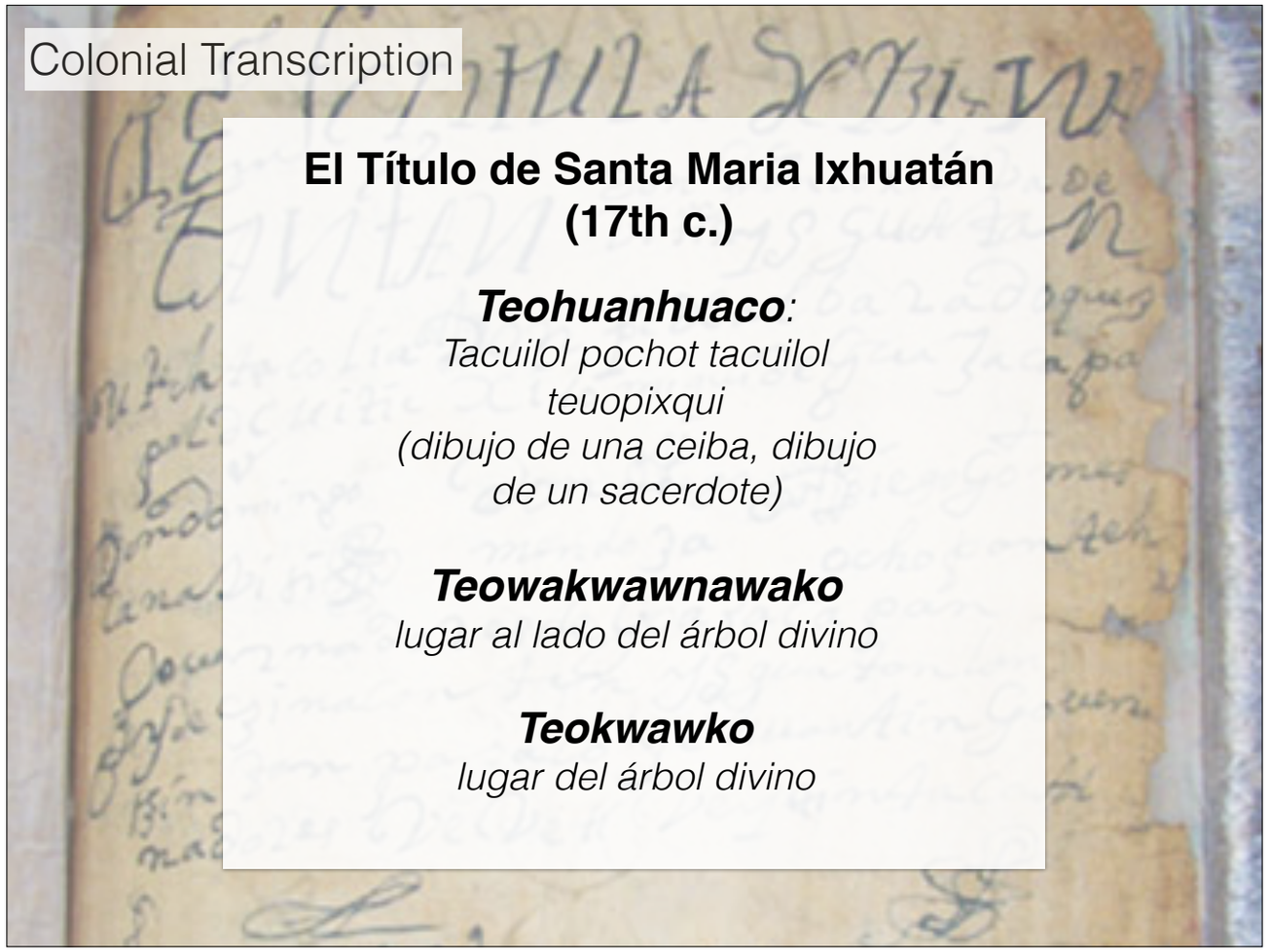
Transcription: Juan González

Our first story is from the sixteenth century. Transcription was everywhere during the beginnings of the Spanish Empire. Notaries transcribed oral testimony, and in doing so, as Kathryn Burns has written, they transformed it by converting it into legal discourse. Sometimes they were even responsible for eliciting the testimony that they transcribed, taking the role of both copyist and compositor.

In New Spain, the conditions of the contact zone complicated matters. Oral testimony, like the huehuetlatolli, is transcribed into alphabetic script. Pictographic documents are copied by trilingual students from various regions and levels of pictographic literacy.

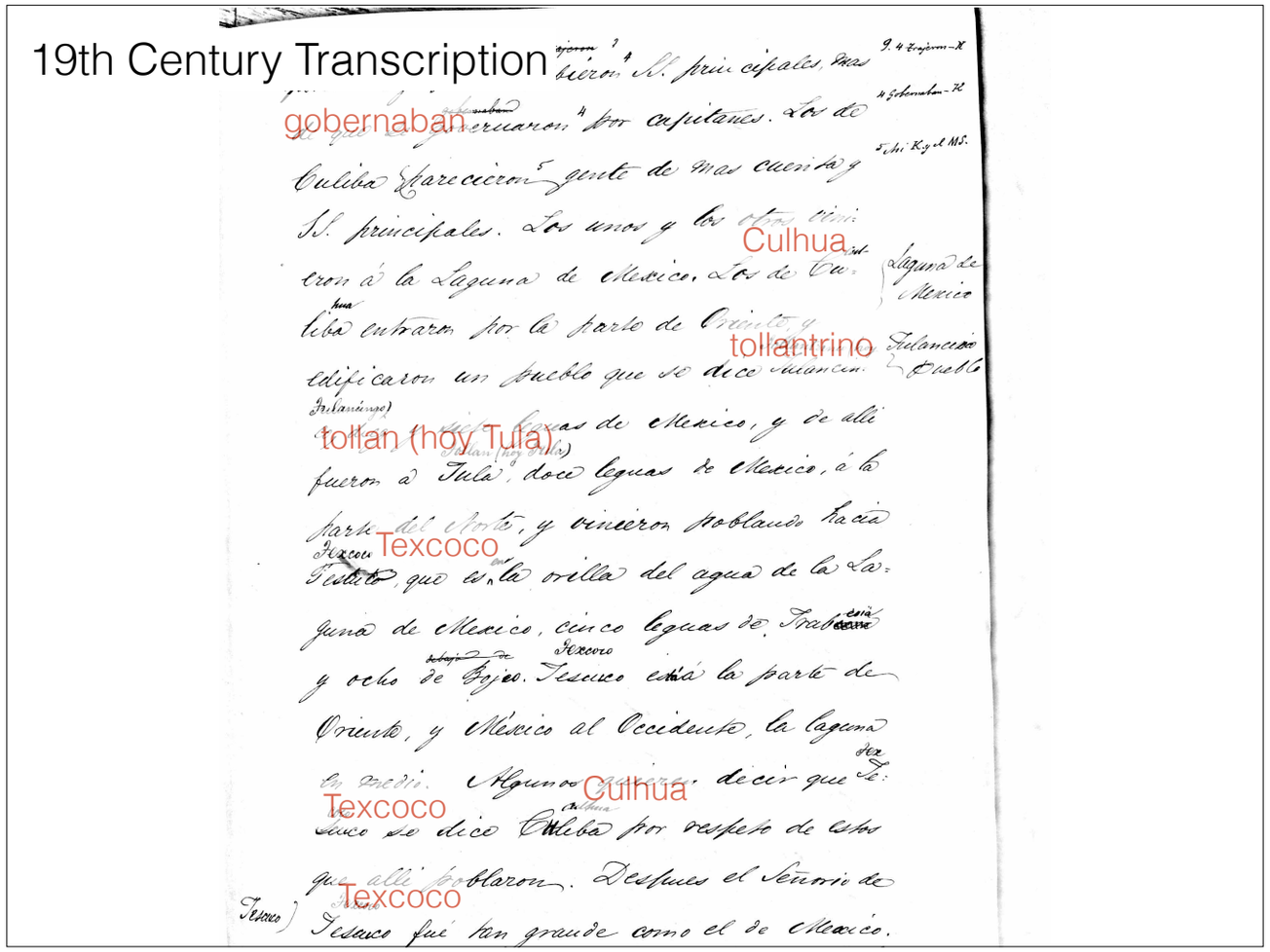


In the florentine Codex, as Ellen Baird has shown, these pictographic transcriptions contain errors that suggest imperfect literacy.



Literacy also impacted the transcription of pictographic documents into alphabetic script. In the case of one seventeenth century *Titulo,* Margarita Cossack Vielman and Sergio Romero have shown how a place name was misrendered into alphabetic nahuatl. It was represented according to a literal reading of the place glyph, rather than a figurative reading.

19th Century Transcription



What about transcription in the nineteenth century? Without going into too much detail, we know that archival records and historical documents circulated in the form of transcribed copies. The result was a transatlantic network of historians, writers, and book collectors who shared documents between Spain, Mexico, and the United States. To give just one small example, the Mexican historian Joaquín García Icazbalceta acquired several of his documents from the U.S. historian William Hickling Prescott, who owned copies from Spain. The document shown here, for example, is a page from Toribio de Benavente Motolinía's Historia. A copy of the document was made in Spain and shipped to the U.S., where Prescott hired a student - often of Italian origin - to make another copy to ship to Icazbalceta. The challenges of paleography, orthography, and language make themselves present on the page. We see dramatic misspellings of Nahuatl words that do not match Motolinía's orthography. We also see frequent spelling errors of Spanish words. We know that Prescott called indigenous Mexican languages "barbarous." Perhaps this is why.

Transcribing Historical Documents:
information loss
changes in signification
changes in value

These stories show how the historical context of transcription can leave its mark on the text and on the way it is received. We've seen how the process of transcription can transform the content of the text, something we would describe today as information loss. We've seen how words can take on different meanings as they are transcribed, and also how a text can signify differently as it passes through different hands.

Reading the First Books

Automatic transcription for multilingual, early modern printed documents

As we consider the automatic transcription of historical printed books (libros antiguos), we should look for the same kinds of things. We should ask how transcription processes leave their mark on the page, and how they change the ways we read words and the ways we think about texts.

With that in mind, we can turn to our 21st-century example, the automatic transcription tools developed for the reading the first books project.

Reading the First Books

Project Team: Sergio Romero, Albert Palacios, Hannah Alpert-Abrams, Maria Victoria Fernández
(UT Austin)

Computer Science: Dan Garrette (U. of Washington), Taylor Berk-Kirkpatrick (UC Berkeley)

Mesoamerican Languages: Stephanie Wood (U. of Oregon), Kelly McDonough, Adam Coon (UT Austin)

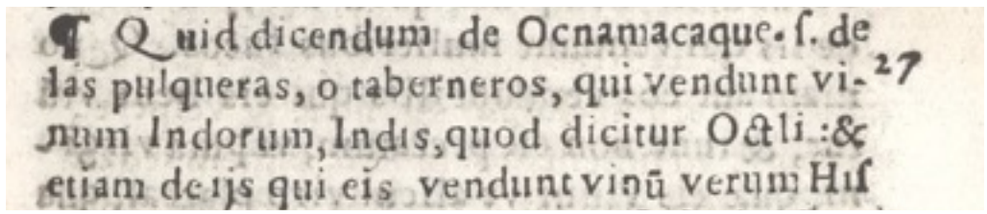
eMOP and Primeros Libros: Laura Mandel, Anton Duplessis, Elizabeth Grumbach (Texas A&M University), Trey Dockendorf, Bryan Tarpley, Matt Christy

University of Texas Libraries: Aaron Choate

The Reading the First Books Project is a collaboration between scholars and librarians at the University of Texas and Texas A&M University.

Multiple Languages

"¶ Q¿nid dicendum de Ocnamacaque. f. de
las pu|queras, o taberneros, qui vendunt vi-
num indorum, indis, quod dicitur Octli .&.
etiam de iis qui eis venduntvr nõ verum Hi|


¶ Quid dicendum de Ocnamacaque. ſ. de
las pulqueras, o taberneros, qui vendunt vi- 27
num Indorum, Indis, quod dicitur Octli :&
etiam de iis qui eis vendunt vinũ verum Hiſ

Bautista, *Advertencias*, 1601

We have been working to develop transcription tools that can handle multiple languages, including indigenous Mesoamerican languages. The tool we developed automatically recognizes and marks shifts between languages, like the ones you see here.

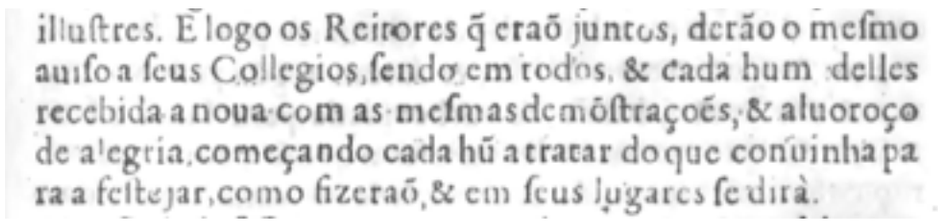
Normalization

Norm.: illustres. E logo os Reitores que taõ juntos, derão o mesmo
Hist.: illustres. E logo os Reitores \~qe taõ juntos, derão o meſmo

Norm.: aviso a seus Collegios, sendo em todos, & cada hum delles
Hist.: aviſo a ſeus Collegios, ſendo em todos, & cada hum delles

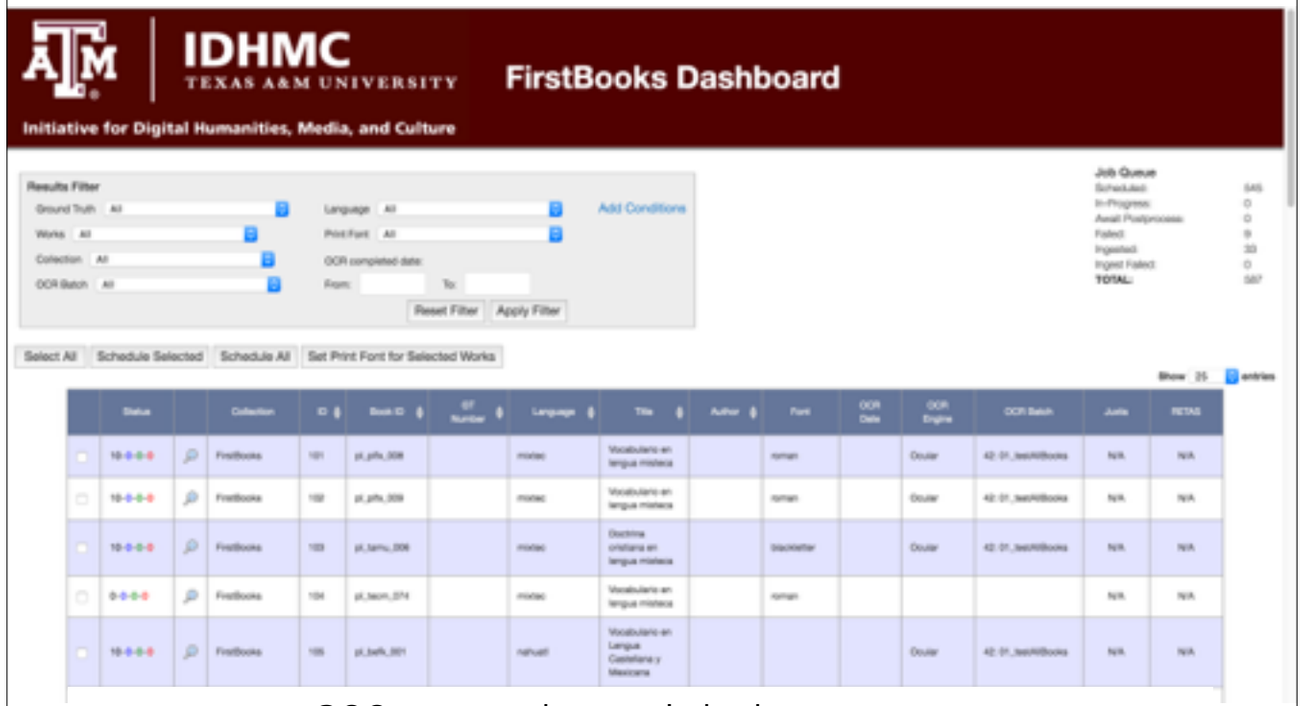
Norm.: recebida a nova com as mesmas demonstrações, & alvoroço
Hist.: recebida a nova com as meſmas demõſtrações, & aluoroço

Norm.: de alegria começando cada huma tratar do que convinha pa-
Hist.: de alegria começando cada hũa tratar do que conuinha pa


illuſtres. E logo os Reitores q̃ eraõ juntos, derão o meſmo
auiſo a ſeus Collegios, ſendo em todos, & cada hum delles
recebida a noua com as meſmas demõſtraçoẽs, & aluoroço
de alegria, começando cada hũ a tratar do que conuinha pa
ra a feſtejar, como fizeraõ, & em ſeus lugares ſe dirà.

We have also extended our transcription tool to automatically learn the difference between historical and modern spelling conventions. So our tool jointly and simultaneously transcribes both a historical (diplomatic) text, and a modern (normalized) text. [This is our first stage results, and we're working on a second stage].

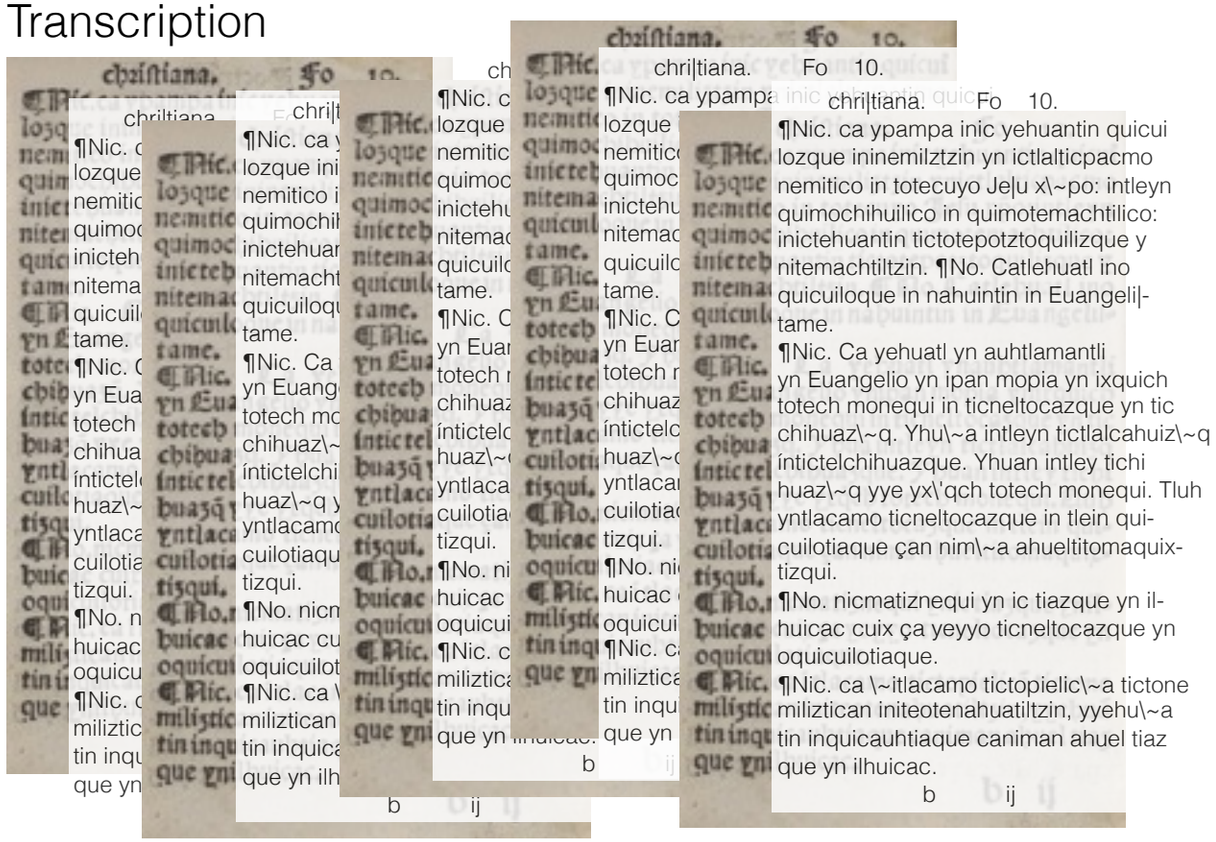
Integrating with eMOP



300 exemplars, eight languages

Finally, we've been working with the Early Modern OCR Project at Texas A&M to build an interface that allows us to use these tools on large corpora of historical texts, such as the sixteenth century printed books in the *Primeros Libros* collection.

Transcription



And this fall, we'll be using that interface to produce transcriptions of all the texts in that collection, which will be made available online. Those transcriptions should create new opportunities for statistical analysis of this collection in the areas of linguistics and book history.

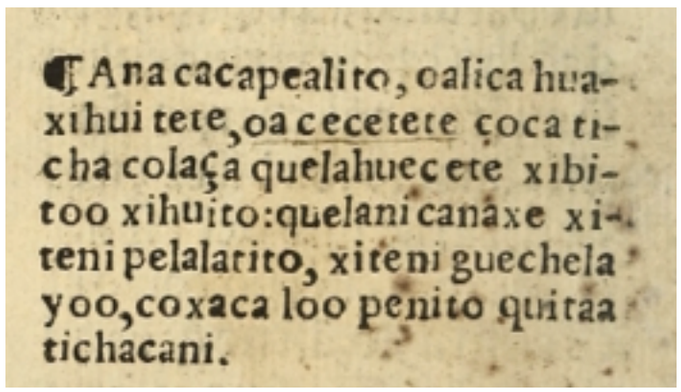
20th Century Transcription

spacing

æ Ana caca pea lito, oalica hua-
xihui teresoa cgçetete goca ti-
cha colaça quelahue cete Xijbi-
too xihui torquelani canâ xe xi-
teni pelala rito, xiteni Sue chela
yoo, coxaca loo pẽiniso guitia
96956894-

overcorrection

dirty OCR



Feria, *Doctrina Cristiana*, 1567

How do the transcriptions of the contact zone, or of transatlantic networks, compare to this digital example? There are a lot of ways we can think about transcription. For example, here you can see one of efforts to transcribe colonial Zapotec. Unfortunately we've found it difficult to find sufficient data to accomplish this kind of transcription. Why is it hard to find data in Zapotec? In part it's because of the history of that language and its presence online, which I think Dr. Lillehaugen can explain more comprehensively than I can. It's also due, in part, to the way that scholars and historians think of transcriptions. Though most scholars of colonial zapotec have folders on their computers filled with transcriptions, they struggle to conceive of those transcriptions as shareable data.

20th Century Transcription

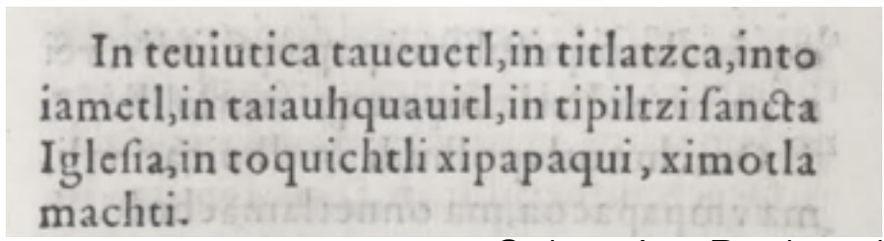
morpheme parsing

teoyotica + tawewetl

Norm.: In teu iutica ta ueuetl, in titlatzca, into-
Hist: In teu iutica ta ueuetl, in titlatzca, into-

Norm: ca metl, in toiauhquauitl, in tepiltzi sancta
Hist: ia metl, in taiauhquauitl, in tipiltzi fancta

Norm: yglesia. in toquichtli xipapaqui, ximotla-
Hist: Iglefia. in toquichtli xipapaqui, ximotla


In teuiutica taucuetl,in titlatzca,into
iametl,in taiauhquauitl,in tipiltzi ſancta
Igleſia,in toquichtli xipapaqui, ximotla
machti.

Sahagún, *Psalmodia*, 1583

Here we can see our efforts to modernize Nahuatl transcription. Like I said, the modernization software is still a work on progress. But you can see here some of the challenges we face…

20th Century Transcription

Norm: lĩ yiollo, oquimopanatili in cemana-
Hist: lĩ yiollo, oquimmopanauili in cemana-

over - correction

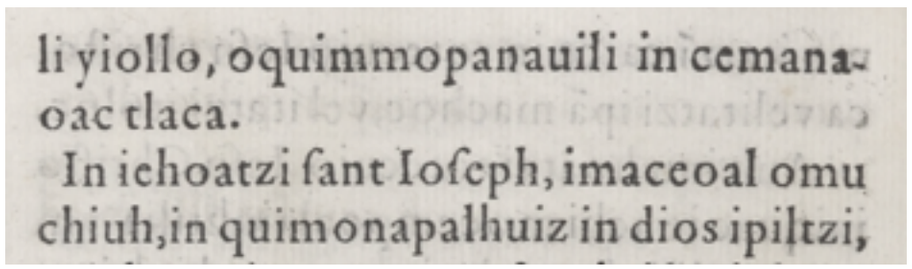
false substitution

Mod: oac tlaca.
Hist: oac tlaca.

Norm: In iehoatzi sant joseph. ymaceoal omo-
Hist: In iehoatzi ſant foſeph. imaceoal omu

correct substitution!

Norm: chiuh, in quimonapalhuiz in dios ypiltzi,
Hist: chiuh, in quimonapalhuiz in dios ipiltzi,


li yiollo, oquimmopanauili in cemana-
oac tlaca.
In iehoatzi ſant Ioſeph, imaceoal omu
chiuh, in quimonapalhuiz in dios ipiltzi,

Sahagún, *Psalmodia*, 1583

… and here are some more. What you see here is that the system has not successfully learned to modernize Nahuatl. This is, again, due to a scarcity of data. It is also due to the complications with the history of Nahuatl orthography standardization. There are competing positions on the way Nahuatl should be modernized that go back centuries, and each document bears that history in its orthography, and in its modernized transcription.

“Reading the First Books”

sites.utexas.edu/firstbooks

Hannah Alpert-Abrams
halperta@gmail.com

To conclude, I hoped to accomplish two goals with this talk. First, I wanted to encourage an approach to thinking about new digital tools within their historical context. Even as we take advantage of the opportunities afforded by these tools, we can consider how contemporary ideologies or the contingencies of access and information shape the work we do.

Second, I wanted to introduce you to the First Books project. All of our tools and transcriptions are online on GitHub. If you are working with a corpus of historical printed documents that you’d like to test with our system, or if you’re interested in learning more about our project, I encourage you to contact us or to visit our website.

Thank you.