

# Transcribing Multilingual and Historical Documents

Hannah Alpert-Abrams  
University of Texas at Austin  
[www.halperta.com/teaching](http://www.halperta.com/teaching)

L L I L A S B E N S O N

# What is transcription?

## Why transcribe?

### Transcribing by machine

### Transcribing by hand

Transcription is:  
the sequential replication  
of text across media.

Recopying is time-consuming, it cramps your shoulder and stiffens your neck. But it is through this action that meaning is discovered.

—Arlette Farge

¶ **A**lic. ca yampamā inic yehuantin quicui  
lozque ininemiliztin ynictlalticpacmo  
neantico in totecuyō Jesu xpō:intleyn  
quimochibuilico in quimotemachtilico:  
inictehuantin tictotepotztoquili que y  
nitemachtiltzin. ¶ **N**o. Catlehuatl ino  
quicuiloque in naḥuaintin in Euangeliſe  
tame.

¶ **A**lic. La yehuatl ynauhlamantli  
yn Euangilio ynapanmopia yniꝝquich  
totech monequi in ticneltocazque yn tic  
chihuazq. Yhuā intleyn tictlalcahuizq  
inticelchihuazque. Yhuan intleytichi  
huazq yee yxqch totech monequi. Auh  
yntlacamo ticneltocazque in tlein qui  
cuilotiaque can nimā ahueltitomaquix  
tizqui.

¶ **N**o. nicmatiznequi ynictiazque yn il  
huicac cuiꝝ ga yeyyo ticneltocazque yn  
oquicuilotiaque.

¶ **A**lic. ca itlacamo tictopielicā tictone  
milizticā in teotenahuatiltzin, y yehua  
tin in quicauhtiaque caniman ahuel tiaz  
que yn ilhuicac.

¶ Nic. ca ypampa inic yehuantin quicui lozque ininemiliztin ynictlaltecpacmo nemitico in totecuyo Jesu xpo: intleyn quimochihuilico in quimotemachtilico: inictehuantin tictotepotztoquilizque y nitemachtitzin. ¶ No. Catlehualt ino quicuiloque in nahuintin in Euangeliſtame.

¶ Nic. Ca yehuatl ynauhtlamantli yn Euangilio yn ipan mopia yn ixquich totech monequi in ticneltocazque yn tic chihuazq. Yhuā intleyn tictlalcahuizq intictelchihuazque. Yhuan intley tichi huazq yye yx'qch totech monequi. Tluh yntlacamo ticneltocazque in tlein qui- cuitotiaque çan nimā ahueltitomaquiax- tizqui.

¶ No. nicmatiznequi yn ic tiazque yn il- huicac cuiçça yeyyo ticneltocazque yn oquicuitotiaque.

¶ Nic. ca itlacamo tictopielicā tictone milizticā initeotenahuatiltzin, yyeahuā tin inquicauhquiaque caniman ahuel tiaz que yn ilhuicac.

¶ Nic. ca ypampa inic yehuantin quicui lozque ininemiliztin ynictlaltecpacmo nemitico in totecuyo Jeſu xpo: intleyn quimochihuilico in quimotemachtilico: inictehuantin tictotepotztoquilizque y nitemachtitzin. ¶ No. Catlehualt ino quicuiloque in nahuintin in Euangeliſtame.

¶ Nic. Ca yehuatl yn auhtlamantli yn Euangilio yn ipan mopia yn ixquich totech monequi in ticneltocazque yn tic chihuazq. Yhuā intleyn tictlalcahuizq intictelchihuazque. Yhuan intley tichi huazq yye yx'qch totech monequi. Tluh yntlacamo ticneltocazque in tlein qui- cuitotiaque çan nimā ahueltitomaquiax- tizqui.

¶ No. nicmatiznequi yn ic tiazque yn il- huicac cuiçça yeyyo ticneltocazque yn oquicuitotiaque.

¶ Nic. ca itlacamo tictopielicā tictone milizticā initeotenahuatiltzin, yyeahuā tin inquicauhquiaque caniman ahuel tiaz que yn ilhuicac.

# Let's practice!

## PIANOS THAT SERVE<sup>‘</sup>



Smithsonian Transcription Center:  
National Baptist Metoka and Galeda Bible Class Magazine, September 1917

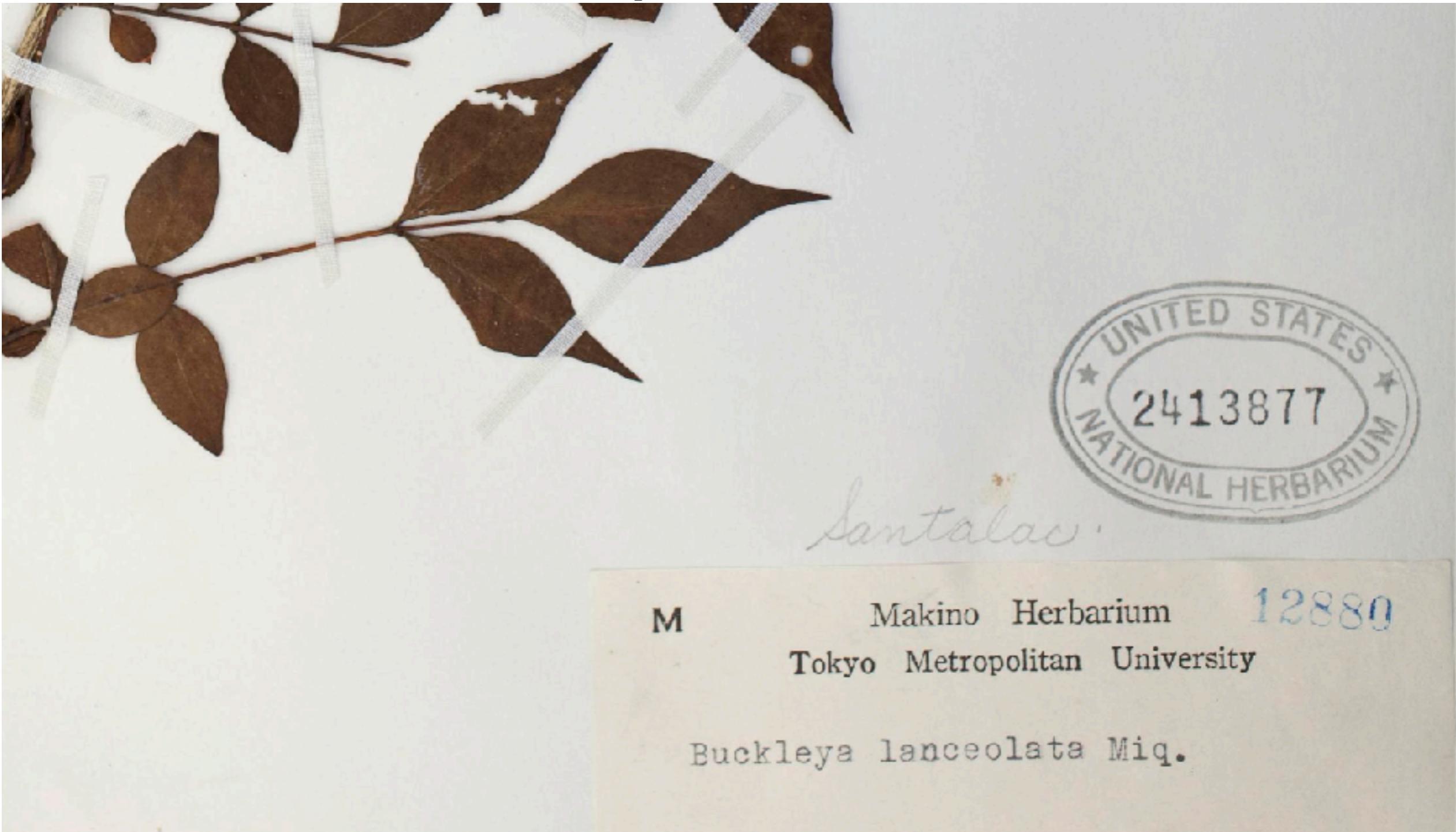
# Let's practice!



[[boxed advertisement]]  
PIANOS THAT SERVE  
[[double line]]  
[[image - small floral sketch]]

Smithsonian Transcription Center:  
National Baptist Metoka and Galeda Bible Class Magazine, September 1917

# Let's practice!



Smithsonian Transcription Center:  
Herbarium Project: Santalaceae

# Let's practice!

**TRANSCRIPTION FORM** INSTRUCTIONS

1) General    2) Collector Details    3) Location

Collector #  Collection Date

Collector

Notes on Transcribing this page (optional)

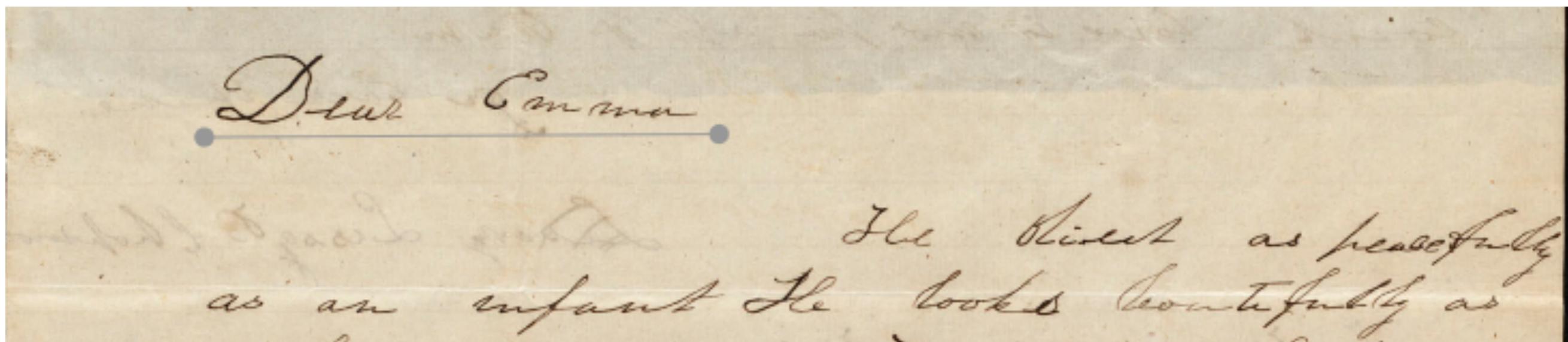
Makino Herbarium 12880  
Tokyo Metropolitan University

Buckleya lanceolata Miq.



Smithsonian Transcription Center:  
Herbarium Project: Santalaceae

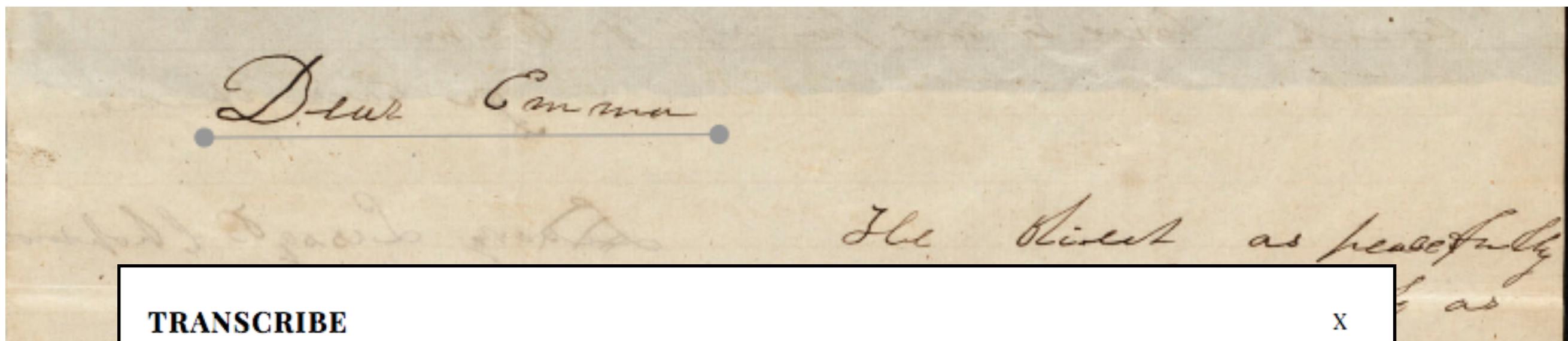
# Let's practice!



Anti-Slavery Manuscripts  
[antislaverymanuscripts.org](http://antislaverymanuscripts.org)

Zooniverse & Boston Public Library

# Let's practice!



## TRANSCRIBE

X

Please transcribe all of the words in the line of text.

TEXT MODIFIERS [insertion] [deletion] [unclear] [underline]

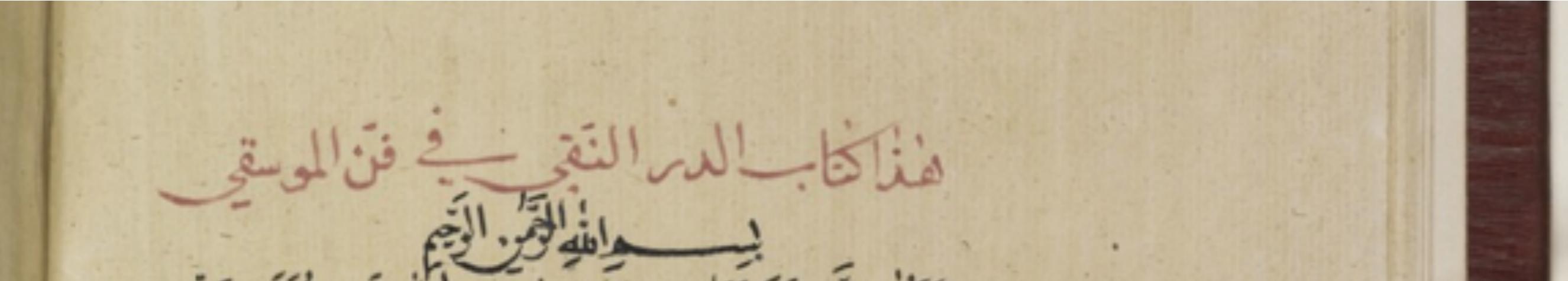
Dear Emma

Delete

Cancel

Done

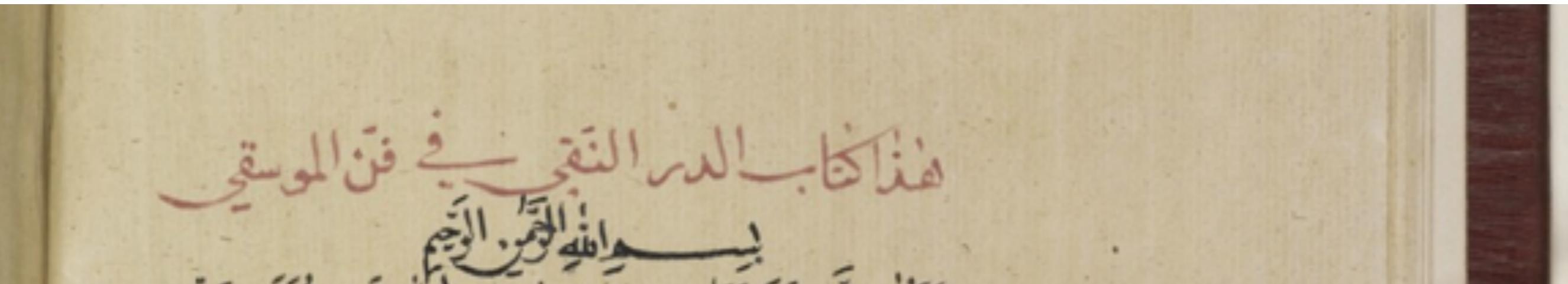
# Let's practice!



هذا كتاب الدر النقي في فن الموسيقى  
بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

al-Durr al-naqī fī fann al-mūsīqī الدّر النّقيّ فِي الْفَنِّ الْمُوسِيقِيِّ Mawsilī, Ahmad ibn ‘Abd al-Rahman  
[2v] (27/70) موصلي، أحمد بن عبد الرحمن

# Let's practice!



هذا كتاب الدر النقي في فن الموسيقى  
بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

- [1] هذا كتاب الدر النقي في فن الموسيقى
- [2] بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

al-Durr al-naqī fī fann al-mūsīqī الدّر النّقيّ فِي الْفَنِّ الْمُوسِيقِيِّ Mawsilī, Ahmad ibn ‘Abd al-Rahman  
[2v] (27/70) موصلي، أحمد بن عبد الرحمن

# Let's practice!



Elizabeth Hill Boone, The Red and the Black

# Let's practice!



Viceroy Mendoza  
(maguey + tozan [gopher])

Elizabeth Hill Boone, The Red and the Black

# What is transcription?

Transcription is interpretation.

# What is transcription?

Transcription is  
specialized interpretive labor.

What is transcription?

Why transcribe?

Transcribing by machine

Transcribing by hand

partida 272<sup>a</sup>

moco de xb. ab o sup. lo  
entrega al alfa  
qui mayor que  
lo fechaba que  
al faghi.



2 gestos los mocos

tlamazqui / los  
alfa / mayz



## The Codex Mendoza

Tlacuilo: Francisco Gualpuyogualcal

Transcription: Juan González

partida  
segunda

viejo

mujer

varon

viejo

padre

hijo

viejo

mujer

varon

viejo

hijo

viejo

mujer

varon

viejo

hijo

gratos que los tenjades guarda  
cos grandes mercedes, y grandes  
beneficios, aueys hecho a este pue-  
blo, yaesta gente quelos aueys  
hablado, como y padres asus  
hijos, aueys hecho el deuer para tain.  
con vuestro pueblo, y los aueys  
declarados, y manifestado los  
secretos de vnas coracones,  
yellor anoydo, y cesidore  
go auestro señor, que lasien  
fan, yentient dan ylopon  
gan por obra, adonde quiera  
que fueren, yestaueren. Plega  
adios, que con la primas se auer  
con este beneficio, y consel se  
consoelen, quando hisieren al  
guna cosa, que no conanenre.  
Señor nuestro, y reynuestro,  
señores senadores, y suzer,  
peruentera ya osdos pena  
con la prolixidad demis pala-  
bras, seors muy bien aventure-  
rados, deos nuestro señor dios  
muchia paß, y asos iego, y viuays  
por muchos años, regiendo, y  
govermando, y gaudando  
auestro señor, con vuestros  
oficios. el qual es invisible, y  
ympalpable.



## The Florentine Codex

Misrendered images reflect  
imperfect literacy



macion mattinemy, Maro  
nijna jntuerny injnerija, my  
quiccalari, auh injuesia a  
ah mae valchocas, maic  
oalmellaquaoas inciquac  
ita iparz choloto, in nom  
lauh, in momotecujn. A  
motzonlecozin amelki  
quijutsin njueoa. Hea  
qujmoma dithia, ma amedz  
mo hamata halili intd. auh  
nacimatlacotilican, maxi  
motequijtilican, maxicmo  
nanam qujlican intloq, in  
naoq que, mioalli, in checatl.

Capitulo desisiete, del rasanamiento, lleno  
demuy buena doctrina enlomoral, que el señor  
hacia sus hijos, quando

ya auja llegado alos a-  
ños de desicion: exor-  
tandolos ahuyr, losvi-  
cios, yaque se diesen,  
los exercicios denoble-  
za, y de virtud.

Iccax tolli omome ca-  
pitulo, vncan motereoaa,  
centhamati circa qualli, le  
nonotsaliztlatelli, necemij  
listiloz, injc qujn nonotsa-  
ja, i pilhoan tlatoanj: inj  
quacie ixtlamati, ietlaca  
que, qujn tlahuauh mo-  
raio, injc qujtlacahujs  
que, injc qujch in aqualli,  
mnaectli. Auh injc qujtlacahujs  
quauhtsitzqujque, impilte  
que, intlatocatequjtl: auh  
injxqujch qualli iectli.

## El Título de Santa María Ixhuatán (17th c.)

### **Teohuanhuaco:**

*Tacuilo pochot tacuilo  
teuopixqui*

*(dibujo de una ceiba, dibujo  
de un sacerdote)*

### **Teowakwawnawako**

*lugar al lado del árbol divino*

### **Teokwawko**

*lugar del árbol divino*

# 19th Century Transcription

bieron<sup>3</sup> S. principales, que  
gobernaban<sup>4</sup> por capitanes. Los de  
Culiba aparecieron<sup>5</sup> gente de mas cuantia y  
S. principales. Los unos y los otros vini-  
eron á la Laguna de Mexico. Los de Culiba entraron por la parte de Oriente, y  
edificaron un pueblo que se dice Culancion<sup>6</sup> Pueblo  
<sup>Culaniungs</sup> Tlanco<sup>7</sup> Tlancio  
tollan (hoy Tula)<sup>8</sup> Tlancio  
fueron a Tula, doce leguas de Mexico, á la  
parte del Norte, y vinieron poblando hacia  
Texcoco Texcoco<sup>9</sup> Texcoco  
Tlancio, que es la orilla del agua de la La-  
guna de Mexico, cinco leguas de Tlancio  
y ocho de Tlancio. Texcoco está la parte de  
Oriente, y Mexico al Occidente, la laguna  
en medio. Algunos quieren decir que Te-  
xcoco se dice Culiba por respeto de estos  
que allí poblaron. Despues el Señorío de  
Texcoco Texcoco<sup>10</sup> Texcoco  
Texcoco fue tan grande como el de Mexico.

# 21st Century Transcription

"¶ Quid dicendum de Ocnamacaque. f. de  
las pulqueras, o taberneros, qui vendunt vi-  
num indorum, indis, quod dicitur Octli .&  
etiam de iis qui eis venduntur non verum His

¶ Quid dicendum de Ocnamacaque. f. de  
las pulqueras, o taberneros, qui vendunt vi-<sup>27</sup>  
num Indorum, Indis, quod dicitur Octli :&  
etiam de iis qui eis venduntur non verum His

# Why transcribe?

Accessibility  
Discoverability  
Preservation  
Research

# Why transcribe?

Labor  
Cost  
Skill  
Accuracy  
Privacy

What is transcription?

Why transcribe?

Transcribing by machine

Transcribing by hand

¶ Nic. ca ypampa inic yehuantin quicui lozque ininemiliztin ynictlalticpacmo nemitico in totecuyo Jesu xpo: intleyn quimochihuilico in quimotemachtilico: inictehuantin tictotepotztoquilizque y nitemachtitzin. ¶ No. Catlehuatl ino quicuiloque in nahuintin in Euangeliſtame.

¶ Nic. Ca yehuatl ynauhtlamantli yn Euangilio yn ipan mopia yn ixquich totech monequi in ticneltocazque yn tic chihuazq. Yhuā intleyn tictlalcahuizq intictelchihuazque. Yhuan intley tichi huazq yye yx'qch totech monequi. Tluh yntlacamo ticneltocazque in tlein qui- cuitotiaque çan nimā ahueltitomaquiax- tizqui.

¶ No. nicmatiznequi yn ic tiazque yn il- huicac cuiçça yeyyo ticneltocazque yn oquicuitotiaque.

¶ Nic. ca itlacamo tictopielicā tictone milizticā initeotenahuatiltzin, yyeahuā tin inquicauhquiaque caniman ahuel tiaz que yn ilhuicac.

¶ Nic. ca ypampa inic yehuantin quicui lozque ininemiliztin ynictlalticpacmo nemitico in totecuyo Jeju x~po: intleyn quimochihuilico in quimotemachtilico: inictehuantin tictotepotztoquilizque y nitemachtitzin. ¶ No. Catlehuatl ino quicuiloque in nahuintin in Euangeliſtame.

¶ Nic. Ca yehuatl yn auhtlamantli yn Euangilio yn ipan mopia yn ixquich totech monequi in ticneltocazque yn tic chihuaz~q. Yhu~a intley tictlalcahuiz~q intictelchihuazque. Yhuan intley tichi huaz~q yye yx'qch totech monequi. Tluh yntlacamo ticneltocazque in tlein qui- cuitotiaque çan nim~a ahueltitomaquiax- tizqui.

¶ No. nicmatiznequi yn ic tiazque yn il- huicac cuiçça yeyyo ticneltocazque yn oquicuitotiaque.

¶ Nic. ca ~itlacamo tictopielic~a tictone milizticā initeotenahuatiltzin, yyeahu~a tin inquicauhquiaque caniman ahuel tiaz que yn ilhuicac.

**C**hic. ca yampampa inic yehuantin quicui lozque ininemiliztin ynictlalticpac mo neantico in totecuyó Jesu xpo:intleyñ quimochibuilico in quimotemachtilico: inictehuantin tictotepotztoquili que y nitemachtiltsin. **C**Mo. Catlehuatl ino quicuiloque in nahtuintin in Euangeliſtame.

**C**hic. La yehuatl ynauhlamantli yn Euangilio ynipan mopia yniꝝquich totech monequi in ticneltocazque yn tic chibuažq. Yhuá intleyñ tictlalcahuizq inticelchibuažque. Yhuán intleyñ tichi huazq yxe yxqch totech monequi. Auh yntlacamo ticneltocazque in tlein qui cuiotiaque can nimā ahueltitomaquiažtizqui.

**C**Mo. nicmatiznequi ynictiazque yn il huicac cuiꝝ ga yeyyo ticneltocazque yn oquicuilotiaque.

**C**hic. ca itlacamo tictopielicā tictone milizticā in teotenahuatiltzin, y yehuá tin in quicauhtiaque caniman ahueltiaz que yn il huicac.

b ij

o — M r^ -^ v^SO |>»00 Cv O — «  
r^ -«t- «^lo r^-oo Oso — r^  
vo SO vO \0 vO \*0 vO vO nO so \0  
sOvO vososOnOvOvOvosOvOvO

00 (7s O — «\* < '♦N ^ w^sO r->00  
(7s o «- n r^ 't- «^VO f>.00 Cv o  
w^ w^vO sovovosovovovo>ovo r\*\*  
r^r^i^i^c^r>.c^r>. r\*\*00

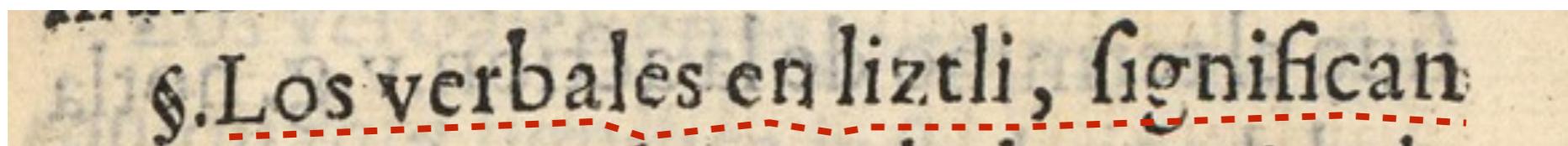
VO r\*\*00 Cv o — \*^ «^ -^ «^SO  
t>i30 J\ o — r«

o — M r^ -^ v^SO |>»00 Cv O — «  
r^ -«t- «^lo r^-oo Oso — r^  
vo SO vO \0 vO \*0 vO vO nO so \0  
sOvO vososOnOvOvOvosOvOvO

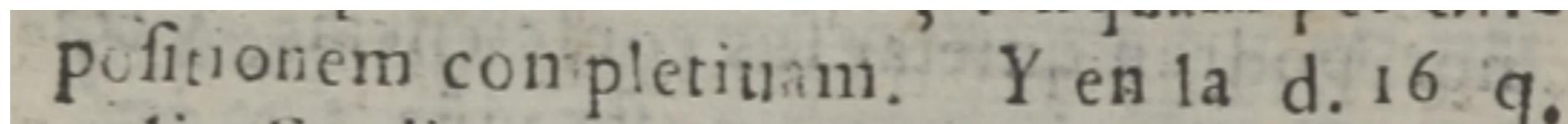
00 (7s O — «\* < '♦N ^ w^sO r->00  
(7s o «- n r^ 't-

# Material Challenges

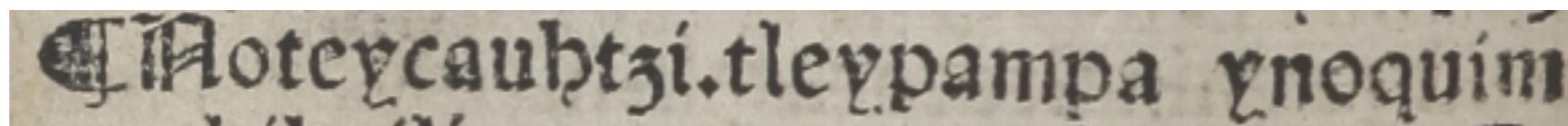
## Wandering baseline



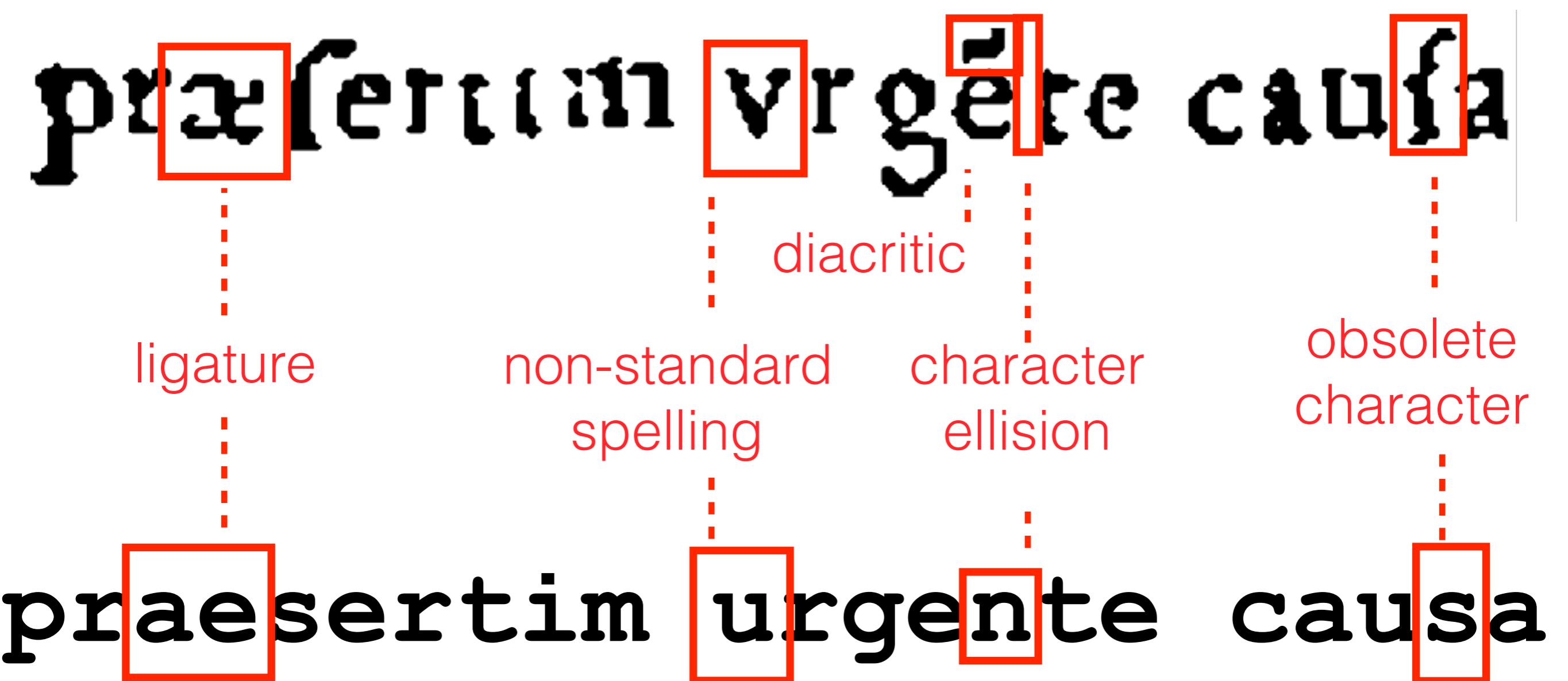
## Uneven inking



## Unfamiliar Typefaces



# Orthographic Challenges



# How does OCR work



|

a o it

a o u

a c u

ai

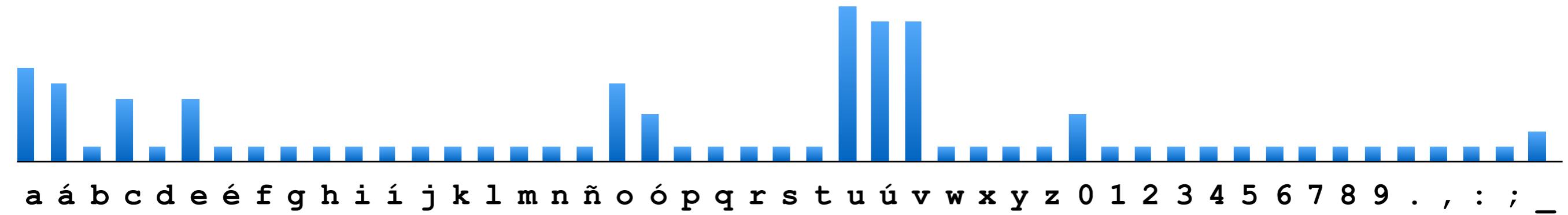
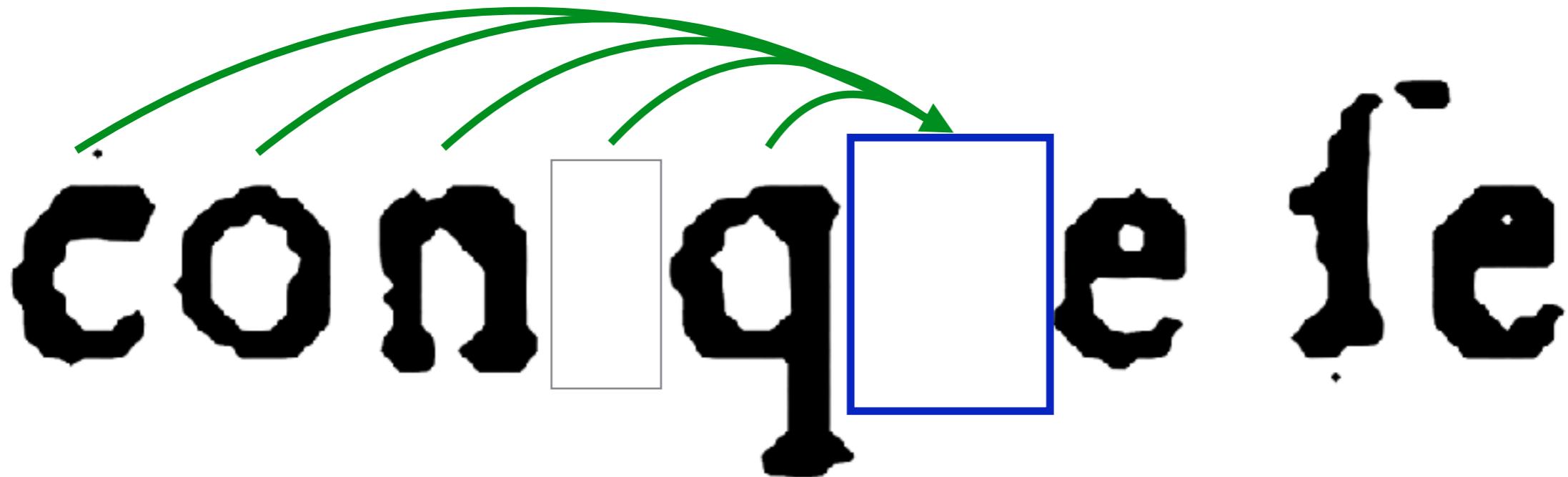
a o ll

a o ll

a c u

ll

a á b c d e é f g h í j k l m n ñ o ó p q r s t u ú v w x y z 0 1 2 3 4 5 6 7 8 9 . , : ;



con que te

con que fe

# Options for Automatic Transcription

Et at id quod obiicitur Clementem Septimum solum fuisse locutum de priuilegiis cōcedendis ordinibus suo tempore a Sede Apostolica approbatis, & religionem S O C I E T A T I S I E S V multo post, fuisse a Sede Apostolica approbatam : responderet, quod post Clementem Septimum, Paulus Quartus, Pius Quartus, & Pius Quintus, & Gregorius Decimtis tertius, & Sextus Quintus priuilegia Mendicantium approbantes, ea omnia de novo concesserunt ; ac si tenor ipsorum priuilegiorum in eorum literis Apostolicis apponetur, & sic priuilegium Clementis Septimi tempore & aetate horum Summorum Pontificum fuisse concessum, quo tempore, & aetate, religio S O C I E T A T I S I E S V non solum erat ab Ecclesia approbata, sed etiam per orbem dispersa. Verum hoc est certum & indubitatum Adrianum Sextum Pontificem maximum ad instantiam Caroli maximi Romanorum Imperatoris & Hispaniarum Regis concessisse fratribus minoribus de obseruancia existentibus, aut ire procurantibus, in di-

Los Con feffores, A Er atid quod obiicitur Clementem Septinit folum fuisse locutum de priuilegiis coceden dis ordinibus suo tempore a Sede Apostolica approbatis, & religionem SOCIETA TISHTES Va multo poft, fuisse a Sede Apostolica approbatam : respondet quod poft Clementem Septimum, Paulus Quartus, Pius Quartus, & Pius Quintus, & Gregorius Decimtis tertius, & Sextus Quintus priuilegia Mendicantium approbantes, ea omnia de Holo concefferunt : ac fi tenor ipforum priuilegiorum in eorum literis Apostolicis apponetur, & sic priuilegium Clementis Septimi tempore & aetate horum Summorum Pontificum fuisse conceffum, quo tempore, & aetate, religio S O C I E T A T I S I E S V. non folyui erat ab Ecclesia approbata, fed etiam Per orbem dispersa. Verum hoc est certum & indubitatum Adrianlin Sexilm Pontifice maximum ad instantiam Caroli maximi Romanorum Imperatoris & Hispaniarum Regis concessiffe fratribus minoribus de obseruancia existentibus, aut ire procurantibus, in die

Free  
Easy  
Limited Quantity  
No training or modification

---

Google Drive

LICENSING FOR ORGANIZATIONS

# ABBYY FineReader 14

Your documents in action.

DOWNLOAD TRIAL

BUY NOW

[Overview](#)

[What's new](#)

[In details](#)

[Why FineReader](#)

[Pricing](#)

FineReader is an all-in-one OCR and PDF software application for increasing business productivity when working with documents. It provides powerful, yet easy-to-use tools to access and modify information locked in paper-based documents and PDFs.



## ABBYY Finereader



Products

Solutions

LICENSING FOR ORGANIZATIONS

## ABBYY FineReader

Your documents in action

[DOWNLOAD TRIAL](#)[BUY NOW](#)[Overview](#)[What's new](#)

FineReader  
productivity  
access



Costly (\$200 - \$600)  
Accurate  
Limited Fonts  
Limited Languages  
Supervised Learning  
Single-Document

## ABBYY Finereader



This organization

Search

Pull requests Issues Gist

▲ + ↻



tesseract-ocr

Repositories

People 0

Search repositories...

Type: All ▾

Language: All ▾

### tesseract

Tesseract Open Source OCR Engine (main repository)

machine-learning ocr tesseract lstm tesseract-ocr

C++ 8,893 2,277 Updated 3 hours ago



### Top languages

C++ HTML

### langdata

Source training data for Tesseract for lots of languages

243 373 Updated on Feb 21



### People

0

This organization has no public members. You must be a member to see who's a part of this organization.

### tessdata

624 318 Updated on Dec 28, 2016



### docs

# Tesseract



This organization

Search

Pull requests Issues Gist

Bell icon + User icon



tes

Repositories

Search repositories...

[tesseract](#)

Tesseract Open Source C

machine-learning OCR

● C++ ★ 8,893 ⚡ 2,27

[langdata](#)

Source training data for T

★ 243 ⚡ 373 Updated

[tessdata](#)

★ 624 ⚡ 318 Updated on Dec 28, 2016

[docs](#)

Open-Access  
Limited Fonts  
Limited Languages  
Modern Orientation  
Command-Line Operations

Language: All ▾

This organization has no public repositories. You must be a member to see them.

# Tesseract

como parece  
Joan, que dice  
in calo Pater  
& hi tres vnu  
loquál deuen  
todas es di  
dader, o re  
ficion, & aña  
sixtin, Tzín,  
logia, duda  
Tepiltzin, S.  
huel nelli teut  
q. d. Dios es  
personas, vn  
cō la qual re  
Tambien se  
In Dio, o S., ca  
sto, q. a huel  
In Dio, ca T  
sto, in ixtzin  
tlahtohuani. Ca inimeixtin  
q. a iceltzin teutl Dio tlahtohu  
huel nelli teutl Dio, q. a iceltzin  
segundo error] çace Dio trans  
cihtotica v. a alios de sus ministros  
cable en si d

## Spirituſca

cō la qual  
también  
in L. ſus, ca

d. I como parece manifiſto en las palabras de Joan, que dice. Tres sunt qui te timoniē dāt. Deus vero in calo Pater, "Verbum, & Spiritus sanctus" cha propoſitiones. & hi tres vnum sunt. 1. Ioann. ultimū int̄ loquál deuen er in truydos y en eñados, que todas tres diuinas personas con vn Deus veritatem amphibodadero; o reformando la obre dicha propoſitione, y añadiendo ella palabra. In huel ixtintzitzin, con que se quita toda amphino-nas, q. a Tettatzi Tepiltzin, Spiritu sancto, ei personas, q. a ixtintzitzin huel nelli teutl Dio in huel imeixtintzitzanco tre personas, vn solo Dio verdadero todas tres. cō la qual reduplicacion se quita toda dubdada dubdada. También se quita con eltas propoſiciones. In DTO S.ca Tettatzi, Tepiltzin, Spiritu sancto. q. a huel iceltzin teutl Dio tlahtohu, Spiritus sancto, in ieixtin personas q. a huel iceltzin tlahtohuani. Ca inimeixtin personas me ca piritus sancto. q. a huel iceltzin teutl Dio tlahtohuani in huel Dio



Ocular

Designed for historical documents  
Handles multiple languages  
Handles nonstandard orthographies  
Automated normalization  
Unsupervised learning

Ocular

# Technologically complex Command-line operations Single-document transcription

Ocular

Intuitive interface  
Parallel computing  
Multiple tools



Ocular + DH Dashboard

Technologically complex  
Command-line operations  
Neural network machine learning  
+ page segmentation

---

Google Vision API

# Reading the First Books

Multilingual, early-modern OCR  
for the Primeros Libros

[sites.utexas.edu/firstbooks](http://sites.utexas.edu/firstbooks)  
[primeroslibros.org](http://primeroslibros.org)  
[emop.tamu.edu](http://emop.tamu.edu)





# The Primeros Libros Project

21 Partner Institutions

430 Exemplars

8 Languages

[www.primeroslibros.org](http://www.primeroslibros.org)

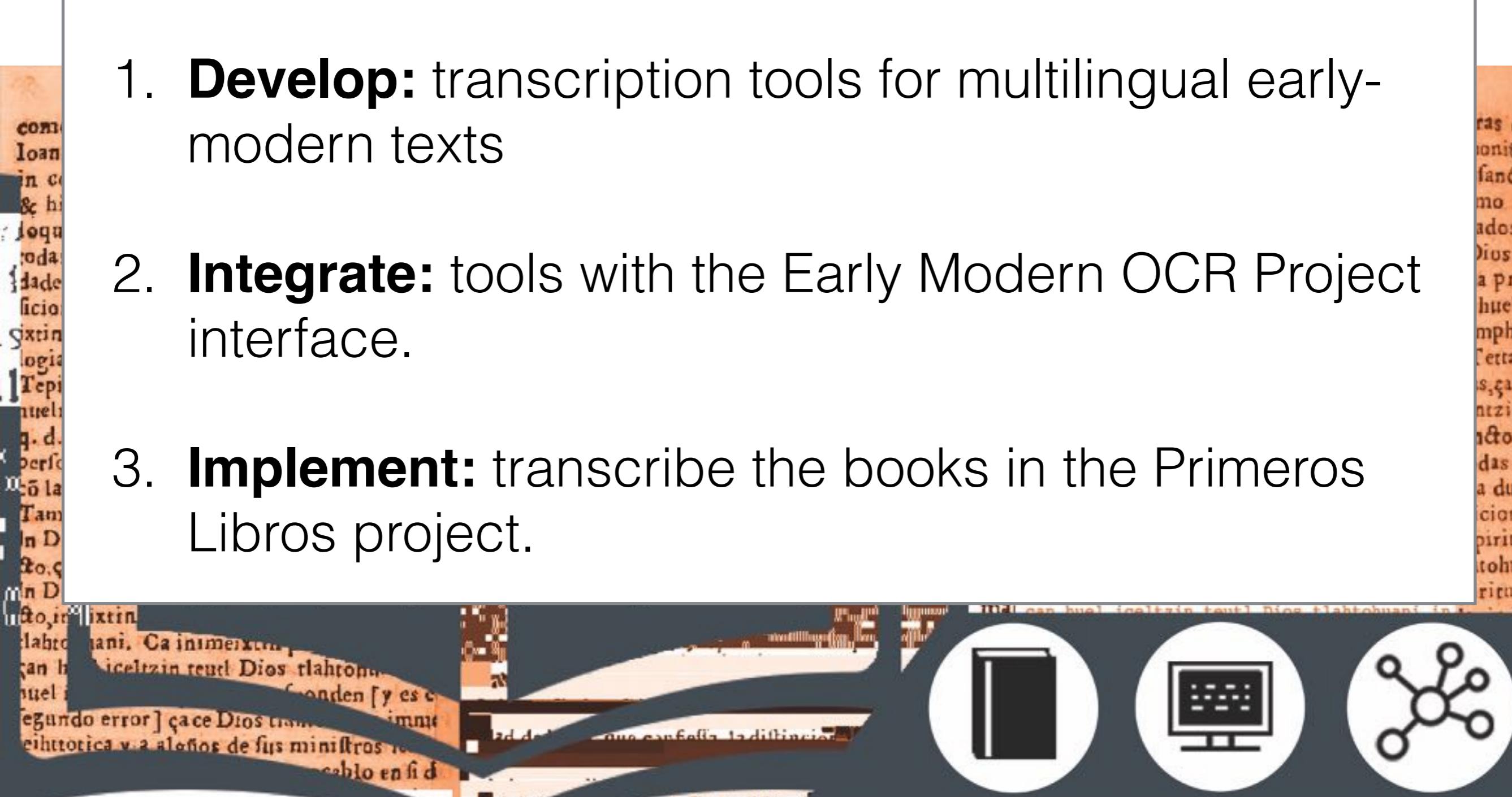
CAPIT. V. VN CAN M  
toa, motenua, in micētlamantli tl  
matiliztli in itechca Anima, iuhqui  
ma micētlamantli xochiqualqua  
yanitl, icenca vel ita

<sup>a quo</sup> tenda.  
C. CAPI. V. CÁYTI.  
mä, y hinéni, nocquencahimaynguëti  
naphäti nocqyquattayxi noccâaia,  
ste ácoh mayguëti etzaten  
tanotzé

# Reading the First Books

## Project Goals:

1. **Develop:** transcription tools for multilingual early-modern texts
2. **Integrate:** tools with the Early Modern OCR Project interface.
3. **Implement:** transcribe the books in the Primeros Libros project.



# Multilingual OCR

"¶ Quid dicendum de Ocnamacaque. f. de  
las pulqueras, o taberneros, qui vendunt vi-  
num indorum, indis, quod dicitur Octli .&  
etiam de iis qui eis venduntvr nō verum His

¶ Quid dicendum de Ocnamacaque. f. de  
las pulqueras, o taberneros, qui vendunt vi-<sup>27</sup>  
num Indorum, Indis, quod dicitur Octli :&  
etiam de ijs qui eis vendunt vinū verum His

# Orthographic Variability

	Original form	Modern form
dize	dize	dice
número	numero	número
Dõde	Dõde	Donde

# Orthographic Variability

Two spellings of ‘mentira’ from one page of one book

mentira

niçira

# Orthographic Variability

merīta

Without handling orth. variation: merita

Correct diplomatic transcription: mētira

Correct normalized form: mentira

# Orthographic Variability

Original image

de las dos que se siguen en las cuales apro  
uecha mucho acostúbrar el anima á se le-

Our diplomatic

de las dos que se siguen en las cuales apro  
uecha mucho acostúbrar el anima á se le-

Our normalized

de las dos que se siguen en las cuales apro  
uecha mucho acostumbrar el ánima á se le-

# Orthographic Variability

Original image  
\_\_\_\_\_

Baseline

Our diplomatic

Our normalized

Gold diplomatic

Gold Normalized

Jefu x̄po
Nefuxpo
*Jefuxpo
*Jesuxpo
Jefu x̄po
Jesu Cristo

# Reading the First Books

## Project Goals:

1. **Develop:** transcription tools for multilingual early-modern texts
2. **Integrate:** tools with the Early Modern OCR Project interface.
3. **Implement:** transcribe the books in the Primeros Libros project.

com  
loan  
in co  
& hi  
joqu  
rada  
dade  
sicio  
sixtin  
logia  
Tepi  
nueli  
q. d.  
perso  
do la  
Tami  
In D  
sto, q  
In D

sto, in sixtin  
tlahtronani. Ca inimerxim,  
gan h uiceltzin teutl Dios tlahtronan  
que uiceltzin teutl Dios tlahtronan  
segundo error] çace Dios tra  
eihitotica v a slos de sus ministros re  
cable en si d



ras de S  
oniâ di  
sanctus  
mo Po  
ados, qu  
Dios ver  
a propo  
huel im  
mphibo  
Cettatzi  
s, çan c  
ntzitzin  
acto tre  
das tres  
a dubda  
ciones.  
piritusa  
tohuani  
ritus an

# Ocular Interface for eMOP

Chrome File Edit View History Bookmarks People Window Help

19% Mon 4:20 PM halperta

DH Dashboard dh-db02.tamu.edu/corpus-manager/ Hannah

All Documents

Show 25 out of 16 rows

Document ID	FB ID	Title	Year	Training Set	View
1059	pl_tamu_015	Advertencias	600	45	Pages Details
1060	pl_tamu_017	Advertencias	600	47	Pages Details
930	pl_tamu_018	Advertencias	600	48	Pages Details
931	pl_tamu_018	Advertencias	600	50	Pages Details
1056	pl_tamu_012	Arte mexicano	595	20	Pages Details
974	pl_tamu_009	Arte y diccionario	574	undefined	Pages Details
<input checked="" type="checkbox"/> 1057	pl_tamu_013	Confesionario	599	43	Pages Details
1058	pl_tamu_014	Contesionario en lengua mexicana y castellana	599	44	Pages Details
1175	pl_tamu_010	De constructione octo partium orationis	579	41	Pages Details
894	pl_tamu_006	Doctrina cristiana en lengua mixteca	568	undefined	Pages Details
1150	pl_tamu_011	Estatutos generales de Barcelona, para la familia cismontana, de la orden de n... u...	585	42	Pages Details
927	pl_tamu_003	Phisica speculativa	557	undefined	Pages Details
928	pl_tamu_004	Reverendi patris fratris Bartholomaei à Ledesma ordinis praedicatorum et sacr...	568	undefined	Pages Details
925	pl_tamu_001	Speculum conjugiorum	558	undefined	Pages Details
926	pl_tamu_002	Speculum conjugiorum	556	undefined	Pages Details
929	pl_tamu_005	Tabula privilegiorum, quae sanctissimus Papa Pius Quintus, concessit fratribus ...	569	undefined	Pages Details
1054	pl_tamu_007	Vocabulario en lengua castellana y mexicana (Volume 1)	571	undefined	Pages Details
1055	pl_tamu_008	Vocabulario en lengua mexicana y castellana (Volume 2)	571	undefined	Pages Details

With Selected: Run a Task

Run a Task

Job Name: pl\_tamu\_013-OCR\_5.29-haa-v03

Job Site: Brazos Supercomputing Cluster

Task: OCR Document with Ocular

Cancel Go

# Ocular Interface for eMOP

Chrome File Edit View History Bookmarks People Window Help

dh-db02.tamu.edu/data/prime

Hannah

DH Dashboard Corpus Manager Job Manager Hello, Hannah! Logout

*Arte mexicana*

**Details**

ID: 1056  
Corpus: FirstBooks  
Author:  
Path: /data/primeros\_llibros/pl\_corpus\_1702/pl\_tamu\_012/1000  
printer: P. Ballí  
ocular\_transcription\_jobs: [{"value": "1094", "name": "pl\_tamu\_012-rincon\_OCR\_5-26-17\_haa"}]  
fb\_work\_id: pl\_tamu\_012  
font\_1: roman  
emop\_work\_id: 165  
ocular\_font\_training\_jobs: [{"name": "pl\_tamu\_012-rincon\_train\_170524", "value": "1088"}, {"value": "1089", "name": "pl\_tamu\_012-rincon\_train\_5-26-17\_haa"}]  
language: nahuatl  
training\_set: 20

**Job Files**

pl\_tamu\_012-rincon\_train\_5-26-17\_haa(1089)  
Ocular Font: 17-05-24\_spalatnah\_500000\_6-v03.fontser  
Ocular Glyph Substitution Model: 17-05-24\_spalatnah\_500000\_6-v03.gemser  
Ocular Language Model: 17-05-24\_spalatnah\_500000\_6-v03.lmser

**Pages**

Page Num: 1  
ID: 134353 Image: pl\_tamu\_012\_00001-1000.jpg  
fb\_page\_id: pl\_tamu\_012\_00001

pl\_tamu\_012-rincon\_OCR\_5-26-17\_haa(1094)  
Ocular Comparison: pl\_tamu\_012\_00001-1000\_comparisons.txt  
Ocular DIPL XML: pl\_tamu\_012\_00001-1000\_diplalto.xml  
Ocular NORM XML: pl\_tamu\_012\_00001-1000\_norm.alto.xml  
Ocular Normalized Transcription: pl\_tamu\_012\_00001-1000\_transcription\_normalized.txt  
Ocular Transcription: pl\_tamu\_012\_00001-1000\_transcription.txt

Page Num: 2  
ID: 134354 Image: pl\_tamu\_012\_00002-1000.jpg  
fb\_page\_id: pl\_tamu\_012\_00002

pl\_tamu\_012-rincon\_OCR\_5-26-17\_haa(1094)  
Ocular Comparison: pl\_tamu\_012\_00002-1000\_comparisons.txt  
Ocular DIPL XML: pl\_tamu\_012\_00002-1000\_diplalto.xml  
Ocular NORM XML: pl\_tamu\_012\_00002-1000\_norm.alto.xml  
Ocular Normalized Transcription: pl\_tamu\_012\_00002-1000\_transcription\_normalized.txt  
Ocular Transcription: pl\_tamu\_012\_00002-1000\_transcription.txt

Page Num: 3

pl\_tamu\_012-rincon\_OCR\_5-26-17\_haa(1094)  
Ocular Comparison: pl\_tamu\_012\_00003-1000\_comparisons.txt

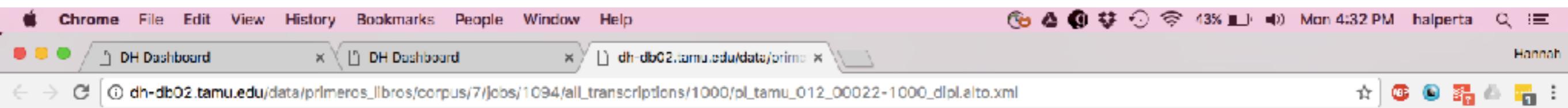
# Ocular Interface for eMOP

The image shows a screenshot of a web browser window with two tabs open. The left tab displays a scanned page from a historical document, showing aged paper with printed Spanish text. The right tab shows the same text in a digital transcription format, likely generated by Optical Character Recognition (OCR) software. Both tabs have the URL [dh-db02.tamu.edu/data/primeros\\_llibros/pl\\_corpus\\_1702/p...](http://dh-db02.tamu.edu/data/primeros_llibros/pl_corpus_1702/p...) and are titled "Hannah". The browser interface includes standard navigation buttons, a search bar, and a status bar at the bottom indicating the date and time.

V. S. en persona los á tomado en si  
porque a llegado V.S. por la vna parte  
hasta la mar del norte, y por la otra ha-  
sta el mar del Sur, q̄ son los vltimos ter-  
minos de su obispado, no perdonado  
qualquier distacia, o aspereza de cami-  
nos, ni a los peligros de los Rios, ni ala  
diuersidad de tantos templos mal fa-  
nos y contrarios a la salud de V.S. an-  
tes lo da todo por bié empleado, por  
cultiuar y beneficiar por sus manos tā  
tas y ta preciosas platas como nuestro  
señor lea encomendado. Por lo qual  
qualquiera ministro se deue cōfundir  
por vna parte de no imitar a quien tie-  
ne obligacion, en padecer algo, y por o  
tra parte se due animar a no huir de este  
pequeño cuidado y sudor que se le pi-  
de en deprender qualquiera legua pa-  
ra abilitarse ē hazer su ministerio. Su  
plico

V. E. en persona los á tomado en si  
porque a llegado v. S. por la vna parte  
hasta la mar del norte, y por la otra ha-  
sta el mar del Sur, o fon los vitimos ter-  
minos de su obispado, no perdonado  
qualquier distacia, o aspereza de cami-  
nos, ni a los peligros de los Rios, ni aja  
diuersidad de tantos templos mal fa-  
nos y contrarios a la falud de V. E. an-  
tes lo da todo por biēemple ado, por  
cultiuar y beneficiar por sus manos tā  
tas y ta preciosas platas como nuestro  
Ieñor lea encomendado. Por lo qual  
qualquiera ministro se deue cōfundir  
por vna parte de no imitar a quien tie-  
ne obligacion, en padecer algo, y por ó  
tra parte se due animara no huirde fle  
pequeño cuidado y sudor que fe le pī  
de en de prender qualquieta legua Pa  
ra abilitarse ē hazer firministerio. Su p-  
plicó

# Ocular Interface for eMOP



This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<alto xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance' xmlns:xlink='http://www.w3.org/1999/xlink' xmlns='http://www.loc.gov/standards/alto/ns-v3#'
  xmlns:emop='http://emop.tamu.edu' xsi:schemaLocation='http://www.loc.gov/standards/alto/ns-v3# http://www.loc.gov/standards/alto/v3/alto.xsd'>
  <Description>
    <MeasurementUnit>pixel</MeasurementUnit>
    <sourceImageInformation>
      <fileName>pl_tamu_012_00022-1000.jpg</fileName>
    </sourceImageInformation>
    <OCRProcessing ID='Ocular0.0.3'>
      <preProcessingStep/>
      <ocrProcessingStep>
        <processingDateTime>2017-05-26T01:08:38</processingDateTime>
        <processingStepSettings>
          -outputPath /fdata/idhmc/firstbooks-input/corpus/7/jobs/1094 -inputDocListPath /fdata/idhmc/firstbooks-input/corpus/7/jobs/1094/t5j1094p22_document_list.txt -inputFontPath
          /fdata/idhmc/firstbooks-input/corpus/7/jobs/1089/17-05-24_spalatnah_500000_6-v03.fontser -inputImagePath /fdata/idhmc/firstbooks-input/corpus/7/jobs/1089/17-05-
          24_spalatnah_500000_6-v03.lsser -inputGsmPath /fdata/idhmc/firstbooks-input/corpus/7/jobs/1089/17-05-24_spalatnah_500000_6-v03.qsser -allowGlyphSubstitution true -updateLM
          false -updateGsm false -emissionEngine DEFAULT -outputFormat dipl,norm,alto
        </processingStepSettings>
      <processingSoftware>
        <softwareCreator>
          Taylor Berg-Kirkpatrick, Greg Durrett, Dan Klein, Dan Garrette, Hannah Alpert-Abrams
        </softwareCreator>
        <softwareName>Ocular</softwareName>
        <softwareVersion>0.0.3</softwareVersion>
      </processingSoftware>
      <ocrProcessingStep>
        <OCRProcessing>
      </OCRProcessing>
    </Description>
    <Layout>
      <Page ID='pl_tamu_012_00022' PHYSICAL_IDN_NR='00022'>
        <PrintSpace>
          <TextBlock ID='par_1'>
            <Textline ID='line_1'>
              <String ID='word_0' WIDTH='21' CONTENT='V' LANG='spa'/>
              <String ID='word_1' WIDTH='5' CONTENT='.' LANG='spa'/>
              <SP WIDTH='11'/>
              <String ID='word_2' WIDTH='12' CONTENT='E' LANG='spa'/>
              <String ID='word_3' WIDTH='6' CONTENT='.' LANG='spa'/>
              <SP WIDTH='9'/>
              <String ID='word_4' WIDTH='25' CONTENT='en' LANG='spa'/>
              <SP WIDTH='11'/>
              <String ID='word_5' WIDTH='83' CONTENT='persona' LANG='spa'>
                <ALTERNATIVE PURPOSE='Normalization'>persona</ALTERNATIVE>
              </String>
              <SP WIDTH='11'/>
              <String ID='word_6' WIDTH='29' CONTENT='los' LANG='spa'/>
              <SP WIDTH='9'/>
              <String ID='word_7' WIDTH='12' CONTENT='á' LANG='spa'/>
              <SP WIDTH='9'/>
              <String ID='word_8' WIDTH='88' CONTENT='tomado' LANG='spa'/>
              <SP WIDTH='9'/>
            </Textline>
          </TextBlock>
        </PrintSpace>
      </Page>
    </Layout>
  </alto>
```

# Reading the First Books

## Project Goals:

1. **Develop:** transcription tools for multilingual early-modern texts
2. **Integrate:** tools with the Early Modern OCR Project interface.
3. **Implement:** transcribe the books in the Primeros Libros project.

com  
loan  
in co  
& hi  
joqu  
rada  
dade  
sicio  
sixtin  
logia  
Tepi  
nueli  
q. d.  
perso  
do la  
Tami  
In D  
sto, q  
In D  
sto, in sixtin  
lahtohuani. Ca inimerxim  
gan h uiceltzin teutl Dios tlahtron  
quel i s fondon [y es c  
segundo error] çace Dios tra  
eihtotica v a alos de sus ministros ro  
cable en si d



# Transcribed PL Corpus



A ante t, & d. 29

Atronar se la muger, nino,cuitapan mauhtia.  
Atronada muger, mocuitapan mauhtiqui.  
Atronara otro con ruido, nite,nacaztitiza.  
Atronado assi, tlancacaztititztli.

¶ Audiencia delos juezes. tlacacoya,tlatzonte-  
coya, tecutlatoloya, teccalli,  
Audencia hazer, nitecutlatoa,nirla,caqui,  
Auelo o aguelo, colli tecol.  
Aullar, nite,coyota.  
Aullador, tecoyouani.  
Aullido, tecoyoualiztli.  
Aun, noma, oc noma.  
A vna hazerse, ticcemitoa, tictocentequilia.  
A vna partey a otra, necoc, neneoc, necoccampa occa-  
pa ixti,yyuccampaixti,yyuntlapalixti.  
Aun aun, cuixçaocti cuixçac.  
Aun no has vuelto, ayate, cenza ayate, ayalitzitz, tica-  
chi.  
Aunque, ymñanel, ymmanel, ymmaçonel, immaçonel-  
iu,maciui, maçoiui, maçoneliui, aço, yuh.  
Ausentarse, canapa niyah, nino, yeltia, ni, cholo, nino,  
tlatia, anixpa.  
Ausencia, canapa yaliztli, netlatiliztli, ateixpa.  
Ausente, ayac, canapayaqui amo ixpa.  
Autor hzedor dios, techiuani, tepiquini, teyocoyani.  
Autoridad de persona, teixmauhtiliztli, mauizticayotl.  
Autoridad tener de persona, nite, ixmauhri, ni, mauriztic.  
Autorizada persona, teixmauhri, teixmamauhri.  
Autoridad de scriptura, tlaneltilioni, tlatolneltilioni.  
Autorizada escritura, tlaneltililli, tlatollaneltillili.  
Autorizar escritura, ni, tlatolneltilia, na, moxtlatolneltilia  
Autoridad tener para hazer algo, ni, nauatle.  
Auaricia tener, ni, teoyeuacati, ni, tlatlameti, ni, tzotzoca  
teuitzti, ni, tzotzocati, atle niccaualiztlatamati, & permis-  
phoram, atle niquixcaria, aninococontlanii.

G

# Transcribed PL Corpus

?? ??

Aronarfe la muger, ni no, cuitapan mauhtia.  
Atró nada muger, mocuitapan mauhtiqui.  
Atronar á otro con ruido, ni te nacaztitza,  
á tronado allí, tlanacaztitztli,  
á[ Audiencia de los Juezes. tlacaco ya  
coya. tecutlatolya, teccalli,  
Audiencia hazer, ni tecutlatoa. nitla-caqui.  
Aue io o aguelo. colli tecol.  
Aul lar, nite, coycua.  
At illa dor. tecoyouatit-  
Auilido. tecoyoualiztli,  
Aun. Jsoma, oc noma.  
A yn a hazerse, ticcemitoa, tictocentequilia,  
Ayna parte y a otra, necoc, nenecoc, necocca  
pa ixti yyuccampaixti yyuntlapalixti.  
Aun aunçcuix çä ocçc97xçac.  
Aun no. ayamo.  
Aun no ha abuelto, ayateçcen casayateçayaíz  
tli.  
Aun que, yntla nel, ymmanel, ymmaçonel, ii  
iuh maciui, maçoiui, maçoneltin, aço yuh.  
Aufentat fe. canapa niyauh. nino-yeltia, ni-cl  
tlatia. anixpa,  
Aufencia. canapa yaliztli, netlatiliztli, atëixpa,  
Aufente. ayac. canapa yaqui amo ipxa.  
Autor hazedor dios. techiuani. tepi quini, te yc  
Autoridad de perfona, teixmauhtiliztli. mau  
Autoridad tener de perfona, ni tesix mauhti, ni  
Autorizada perfona, teixmauhhti, teixmamauj  
Autoridad de feriptura. tlaneltiloni, tlatolti  
Autorizada escriptura, tlaneltillili, tlatollanel  
Autoriza referiptura, ni-tlatol neltilia. na, moxi  
Autoridad tener para hazer algo, ni nauatile,  
Auaticia tener, ni ste oyeuacatl, nist latjanteti,  
teuitzti-ni, tzotzocati. atle niccaualiztlamati.  
photam. atle niquixcatia. anñnocotontlani.

A ante f, & d. 29

Atronar se la muger, nino.cuitapan mauhtia,  
Atronada muger, mocuitapan mauhtiqui,  
Atronara otro con ruido, nite.nacaztitza,  
Atronado assi, tlanacaztitztli,  
¶ Audiencia delos juezes. tlacacoya.tlatzonte-  
coya. tecutlatolya. teccalli,  
Audiencia hazer, nitecutlatoa.nitla.caqui,  
Auelo o aguelo, colli tecol.  
Aullar. nite,coycua.  
Aullador. tecoyouani.  
Aullido. tecoyoualiztli,  
Aun. noma. oc noma,  
A vna hazerse, ticcemitoa, tictocentequilia,  
A vna partey a otra, necoc, nenecoc necoccampa occ-  
pa ixti. yyuccampaixti. yyuntlapalixti.  
Aun aunç cuixçaoç cuixçac.  
Aun no. ayamo.  
Aun no has vuelto. ayate' cenza ayate' ayaliztitz' tica-  
tli.  
Aunque, ymñanel, ymmanel, ymmaçonel, immaçonel-  
iuh,maciui, maçoiui, maçoneliui, aço.yuh.  
Ausentar se, canapa niyauh. nino,yeltia, ni,choloa. nino,  
tlatia.anixpa.  
Ausencia. canapa yaliztli,netlatiliztli, ateixpa.  
Ausente. ayac,canapayaqui amo ipxa.  
Autor hazedor dios. techiuani.tepiquini, teyecoyani.  
Autoridad de persona. teixmauhtiliztli. mauizticayorl.  
Autoridad tener de persona. nite,ixmauhtri,ni,maurztic.  
Autorizada persona. teixmauhhti.teixmamauhhti.  
Autoridad de scriptura. tlaneltiloni, tlatolneltiloni.  
Autorizada escriptura. tlaneltillili, tlatollaneltilili.  
Autorizar escriptura. ni,tlatolneltilia,na,moxtlatolneltilia  
Autoridad tener para hazer algo, ni.nauatile.  
Auaticia tener, ni,teoyeuacati, ni, tlatlameti. ni, tzotzoca  
teuitzti.ni, tzotzocati, atle niccaualiztlamati. & permira-  
phoram. atle niquixcatia, anñnocotontlani.

G

# Transcribed PL Corpus

**Atronada muger,**    mocuitapan mauhtiqui.

Atró nada muger,  
mocuitapan mauhtiqui.

**Atronará otro con ruido,**    nite,nacaztititza,

Atronar á otro con ruido,  
ni te nacaztititza,

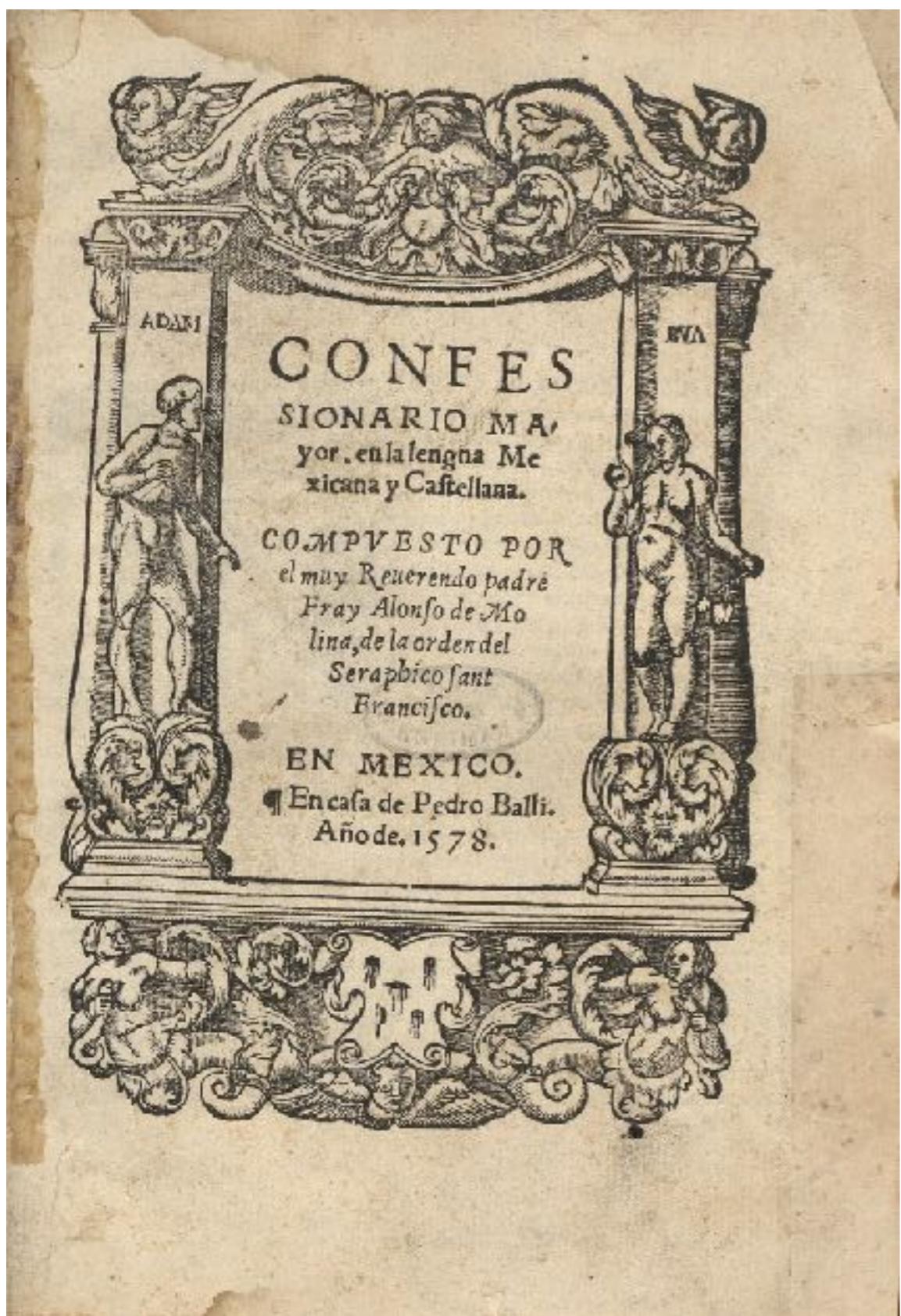
**Atronado así,**    tlanacaztititztli.

á tronado allí,  
tlanacaztititztli,

**■ Audiencia de los juezes.**    tlacacoya

■[ Audiencia de los Juezes.  
tlacaco ya

# Transcribed PL Corpus



Eclengua Mexicana y Castellana. 1578. 10

Ynic amo mopinanbtiz, ma  
nilquitz, inic amo quich-  
teccatlapiquisque, aubcen  
ca momaubtia, quimacaci  
yacuatzacuiltilocatçanno  
yuh motechmonequi, ynic  
cencatimomociniz, ipquich  
motlapal ticchiuaz inicmu  
chi tictemoz, tiquilnami-  
quizimotlatlacol, inic vel  
timogolcuitiz, inic titlapo  
naz yxpantzincototecuixyo  
dios, guaninixpá sacerdo  
te, inic amo tipinauhtiloz,  
inic amo titlatzacuiltyloz,  
çantinaniçonaz ywan ric  
maquixtizimanimá. Ah  
intlacamo achtopa, tiquil-  
namiñiz, riccentlaliz imo  
tlatlacol, vel nelli yctiqui  
tlacozi monegolmelaua  
li, amomanelli yeticpale  
uiñimanimá, çan occenca  
ye ictictoliniz, amoma mo-  
paleuilocamochiuaz imo  
neyolcuitiz, çan mopina  
ubtloca, morelchiualoca  
yez, yuáemicacmitnaua  
tiloca mochiuaz. Ah inic  
amotiquitlacoz, mamocen  
yollocopa achto xictemo,  
guan riccentlali ypixquich  
náten qdioruincueta desí  
y por no ser afretado nica  
er en algúia falta, y por qno  
le tégá por ladron; y tiene  
grámedo y temor del casti  
go que se le podria dar; así  
tienes necesidad de tener  
grá solicitud y cuidado, y  
obazer toda tu posibilidad  
para buscar y pésar todos  
tus pedos para te confessar d  
uidamente, y para dar buéa  
cuéta dlate nro señor dios  
y dlate el sacerdote, porq  
no seas auergoçado y casti  
gado, mas átes recibas ho  
ra, y alcáces la salvació de  
tu alia. Y si pmero no truje  
resala memoria tus pedos  
y los recogieres, ciertame  
te sera tu cōfessió invalida,  
falta y no veradadera, ni me  
nos sera útil y puecposa a  
tu alia, átes la aligiras mis  
chomas, y nosera para tu  
fauor la tal cōfession mas  
para tu cōfession, y cōdena  
ció y para q eternamente se  
as desechado. Y porq nosea  
íperfecta la tal cōfessió en  
ya, busca pmero d todo tu  
corazón aynta en otros pedos  
B 2 tus

# Transcribed PL Corpus

**ynic amo mopīnāuhtij, ma nāten q̄dior uincuētaoe-lī**

ynic amo mopīnāuhtij, ma nāten. q̄Dior uincuētaoe-lī

**uilquirtij, inic amo quicb - y poz no fer atrétado nica**

uilquirtij, inic amo quicb - y poz no fer atrétado nica

**teccatlapiquijque, auh cen eren algūa falta, y poz q̄n.o**

teccatlapiquijque, auh cen

eren algua falta, y poz q̄n.o

**ca momaubtia, quimacaci le tēgā poz ladron ; ytiene**

ca momaubtia, quimacaci

le tēgā poz ladron ; ytiene

# Evaluating OCR

## Google

tlalili. Clic, La epanipa yniquíntech mircuitizqueinqualtin pectin initetlaye colticahuan in sanctome. Ynic mocētla lisque ynic quimotla~~ge~~coltilisque ento tecuiyo:ynarcan iniuhtiquimitta yn teo pirque ynocmigec tlamantin/~~z~~namoti

## Ocular

lalili. di. il lic, ~~TC~~a ypampa yniquíntech mixcuitizque in qualtin yēctin in ītētlaye colticahuan in iancto me. Ynic mocētla lizque ynic quimotlayecoltilizque yn to tēcuiyo: yn axcan in iuh tiquimitta yn teo pixque yn oc miyec tlamantin ? yn amo ti

**tlalili. Clic, La ypampa yniquíntech  
mixcuitizqueinqualtin yectin initetlaye  
colticahuan in sanctome. Ynic mocētla  
lizque ynic quimotla~~ge~~coltilizque yn to  
tecuiyo:ynarcan iniuhtiquimitta yn teo  
pirque ynocmigec tlamantin/~~z~~namoti**

# Evaluating OCR

Evaluating OCR

[www.halperta.com/firstbooks/](http://www.halperta.com/firstbooks/)



# Evaluating OCR

Word Error Rate  
Character Error Rate  
Error Distribution

Gold Standard  
Dictionary  
Language Model



What is transcription?

Why transcribe?

Transcribing by machine

Transcribing by hand

Zooniverse  
Scripto  
From the Page

## What is FromThePage?

FromThePage is software that allows volunteers to transcribe handwritten documents online. This particular instance has been installed by the UT Library to explore opportunities for crowdsourcing and student participation in document transcription projects involving UT collections. Currently it is serving as a platform for a pilot program to transcribe portions of the archives of Texas lawmaker and UT founder [Judge Alexander Watkins Terrell](#), held at UT's [Driscoe Center for American History](#). The papers being transcribed are all connected with Terrell's time as the United States Minister Plenipotentiary to the Ottoman Porte in Constantinople from 1893 to 1897. Initial transcription contributions can be made to the [transcription project of Frank Calvert's letter to Terrell in November of 1894](#).

# Exercise: Transcribing Colonial Documents with From the Page

The FromThePage software is still under development, but we'd like to invite people to look around and send suggestions and bug reports to [benwbrum@gmail.com](mailto:benwbrum@gmail.com). If anything looks broken, hard to understand, or just odd, please let us know! For behind-the-scenes look at the development effort, check out the [product development blog](#).

If you're interested in using FromThePage to host a transcription project, we're looking for you. The software is free to use. Please email [benwbrum@gmail.com](mailto:benwbrum@gmail.com) and tell us about your project.

## Read Transcriptions

Some of Terrell's papers related to his visit to Troy have already been transcribed. Below, you can find Frank Calvert's letter written to Terrell in November of 1894, after Terrell's visit to the site. In this letter, Calvert offers to sell his land at the site to Terrell, and notes that he's appraising the value of the collection.

[Read the Calvert letter](#)

<http://fromthepage.lib.utexas.edu/>

## Transcribe Manuscripts

Work on the transcription of Terrell's papers is ongoing. If you're a student in UCS302: Tales of the Trojan War, fall 2015, you'll be helping with this! To get started, you'll need to [create an account](#). Let your instructor know your username, and he'll give you access to the collection (once you've found and photographed the original documents at the Driscoe Center). Then the collection "Terrell, Calvert and Troy (UCS302)" will appear on your dashboard. Go to it, find a document that hasn't been completely transcribed yet, choose a page, and click the "transcribe" tab. For instructions on transcription conventions and codes for basic formatting (underline, bold, strikethrough, etc.), see the "help" tab on the webpage for a work, or check the user guide in the course site on Canvas.

## Collections

Joaquín García Icazbalceta Collection  
Lista de los pueblos de indios...

## What is FromThePage?

[https://fromthepage.lib.utexas.edu/display/read\\_work?document\\_set\\_id=8&work\\_id=564](https://fromthepage.lib.utexas.edu/display/read_work?document_set_id=8&work_id=564)

FromThePage is software that allows volunteers to transcribe handwritten documents online. This particular instance has been installed by the UT Library to explore opportunities for crowdsourcing and student participation in document transcription projects involving UT collections. Currently, it is serving as a platform for a pilot program to transcribe portions of the archives of Frank Calvert and Terrell, United States Minister Plenipotentiary to the Ottoman Porte in Constantinople from 1893 to 1897. Initial transcription contributions focus on his visit to the archaeological site of Troy in 1894 and his attempts to convince the University to purchase the land and/or collection of Frank Calvert.

The FromThePage software is still under development, but we'd like to invite people to look around and send suggestions and bug reports to [benwbrum@gmail.com](mailto:benwbrum@gmail.com). If anything looks broken, hard to understand, or just odd, please let us know! To stay up-to-date with the development effort, check out the [product development blog](#).

If you're interested in using FromThePage to host a transcription project, we're looking for you. The software is free to use. Please email [benwbrum@gmail.com](mailto:benwbrum@gmail.com) and tell us about your project.

## Read Transcriptions

Cities:

**[[cityname]]**

[Read the Calvert letter](#)

Example:

**[[Oaxaca]]**

## Transcribe Manuscripts

Work on the transcription of Terrell's papers is ongoing. If you're a student in UCS302: Tales of the Trojan War, fall 2015, you'll be helping with this! To get started, you'll need to [create an account](#). Let your instructor know your username, and he'll give you access to the collection (once you've found and photographed the original documents at the Briscoe Center). Once you log in, search for "Terrell, Calvert and Troy" (or whatever name you will apply to your collection). Once you find a document that hasn't been completely transcribed yet, choose a page, and click the "transcribe" tab. For instructions on transcription conventions and codes for basic formatting (underline, bold, strikethrough, etc.), see the "help" tab on the work page for a work, or check the user guide in the [source code repository](#).

People:

**[[last name, first name | name]]**

Ex:

**[[Rojas, Isabel |Isabel de Rojas]]**

What is transcription?

Why transcribe?

Transcribing by machine

Transcribing by hand

# Conclusions

## **Crowdsourcing**

From the Page  
Zooniverse  
Scripto

## **Automatic Transcription**

Google Drive  
Tesseract  
ABBYY FineReader  
Transkribus  
Ocular

# Conclusions

Transcription is interpretive

Transcription is labor

Corpus creation will shape corpus analytics