Transcribing Multilingual and Historical Documents

Hannah Alpert-Abrams
University of Texas at Austin
www.halperta.com/teaching

LLILASBENSON

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

¶ Nic.ca ypampa inic yehuantin quicui lozque ininemiliztzin ynictlalticpac mo nemitico in totecuyo Jesu xpo:intleyn quimochihuilico in quimotemachtilico: inictehuantin tictotepotztoquilizque y nitemachtiltzin. ¶ No. Catlehuatl ino quicuiloque in nahuintin in Euangelis-tame.

¶ Nic. Ca yehuatl ynauhtlamantli yn Euangelio ynipan mopia ynixquich totech monequi in ticneltocazque yn tic chihuazq̃. Yhuã intleyn tictlalcahuizq̃ intictelchihuazque. Yhuan intley tichi huazq̃ yye yxc̃h totech monequi. Auh yntlacamo ticneltocazque in tlein qui-cuilotiaque çan nimã ahueltitomaquix-tizqui.

¶ No. nicmatiznequi ynictiazque ynil-huicac cuix ça yeyyo ticneltocazque yn oquicuilotiaque.

¶ Nic. ca itlacamo tictopielicã tictone miliztican initeotenahuatiltzin, yyehuã tin inquicauhtiaque caniman ahuel tiaz que ynilhuicac.

¶Nic.ca ypampa inic yehuantin quicui lozque ininemilizt̃in ynictlalticpac mo nemitico in totecuyo Jeſu xp̃o:intleyn quimochihuilico in quimotemachtilico: inictehuantin tictotepotzt̃oquilizque y nitemachtiltzin. ¶No. Catlehuatl ino quicuiloque in nahuintin in Euangeliſ- tame.

¶Nic. Ca yehuatl ynauhtlamantli yn Euangelio ynipan mopia ynixquich totech monequi in ticneltocazque yn tic chihuaz̃q. Yhuã intleyn tictlalcahuiz̃q ínticteIchihuazque. Yhuan intley tichi huaz̃q yye yx̊qch totech monequi. Tluh yntlacamo ticneltocazque in tlein qui- cuilotiaque çan nimã ahueltitomaquix- tizqui.

¶No. nicmatiznequi ynic tiazque ynil- huicac cuix ça yeyyo ticneltocazque yn oquicuilotiaque.

¶Nic. ca îtlacamo tictopielicã tictone miliztican initeotenahuatiltzin, yyehuã tin inquicauhtiaque caniman ahuel tiaz que ynilhuicac.

¶Nic. ca ypampa inic yehuantin quicui lozque ininemilztzin yn ictlalticpacmo nemitico in totecuyo Je|u x\~po: intleyn quimochihuilico in quimotemachtilico: inictehuantin tictotepotztoquilizque y nitemachtiltzin. ¶No. Catlehuatl ino quicuiloque in nahuintin in Euangeli|- tame.

¶Nic. Ca yehuatl yn auhtlamantli yn Euangelio yn ipan mopia yn ixquich totech monequi in ticneltocazque yn tic chihuaz\~q. Yhu\~a intleyn tictlalcahuiz\~q ínticteIchihuazque. Yhuan intley tichi huaz\~q yye yx\'qch totech monequi. Tluh yntlacamo ticneltocazque in tlein qui- cuilotiaque çan nim\~a ahueltitomaquix- tizqui.

¶No. nicmatiznequi yn ic tiazque yn il- huicac cuix ça yeyyo ticneltocazque yn oquicuilotiaque.

¶Nic. ca \~itlacamo tictopielic\~a tictone miliztican initeotenahuatiltzin, yyehu\~a tin inquicauhtiaque caniman ahuel tiaz que yn ilhuicac.

Let’s practice!


PIANOS THAT SERVE

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

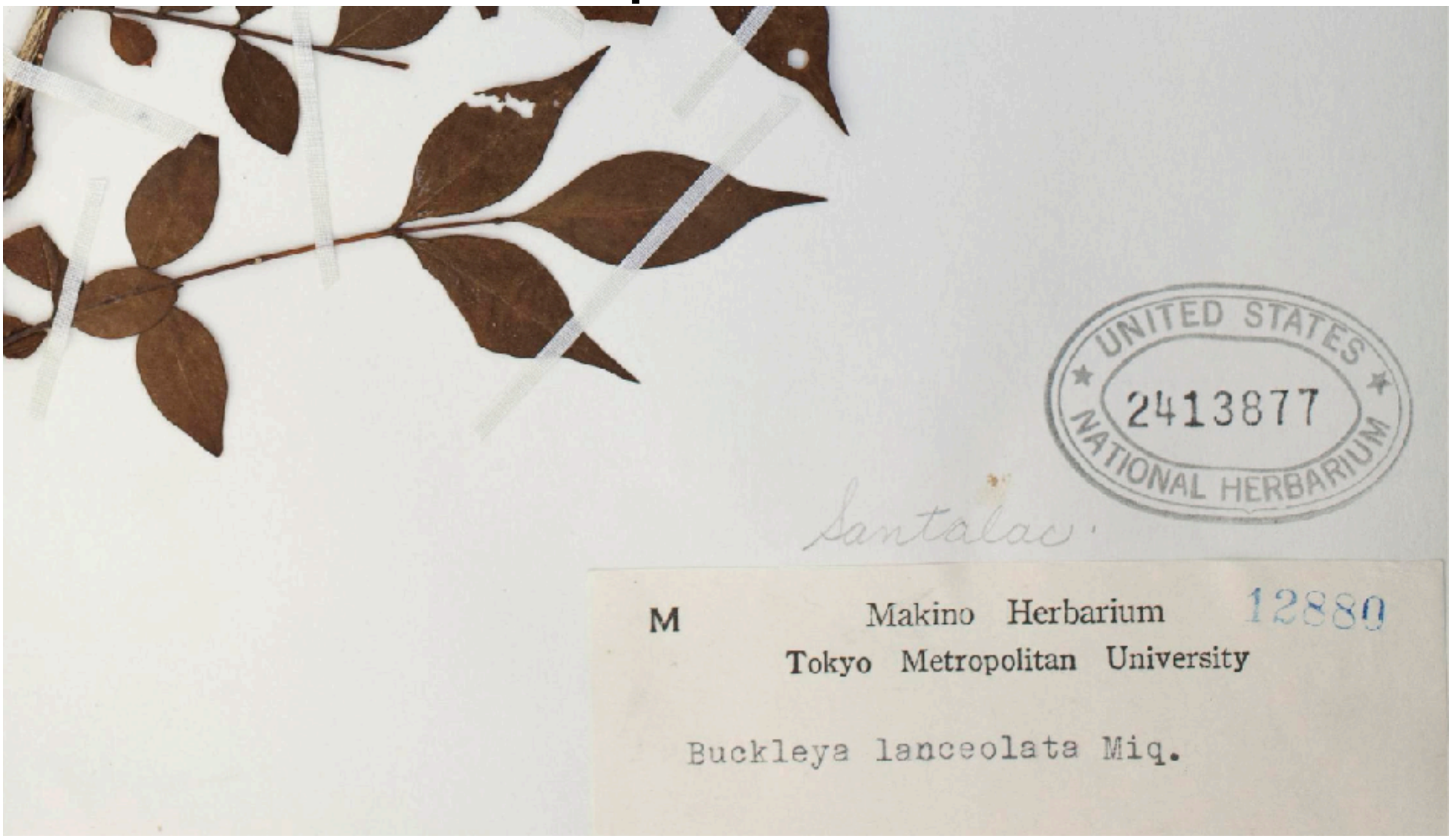
Let’s practice!


PIANOS THAT SERVE

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

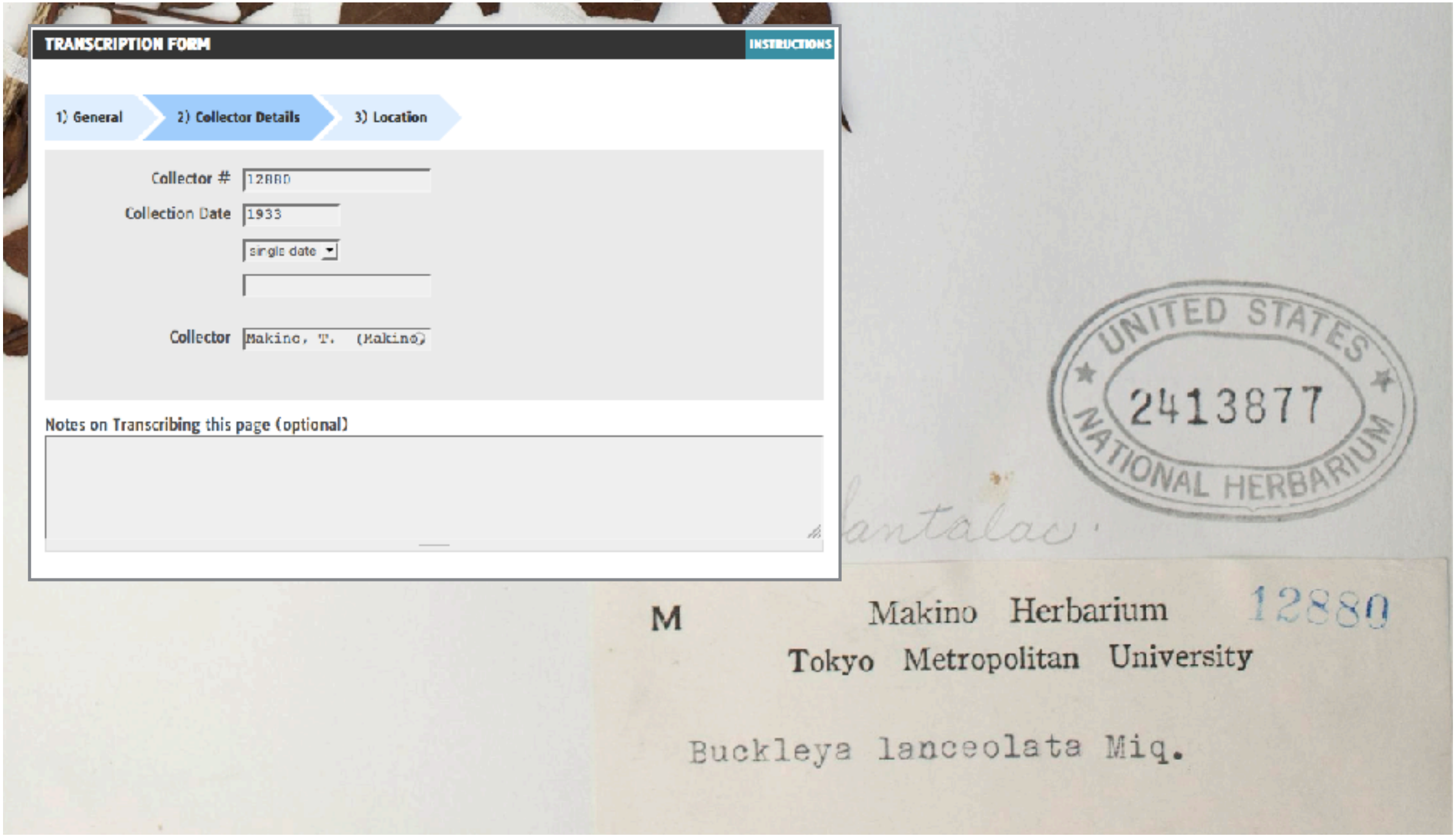
Smithsonian Transcription Center:
National Baptist Metoka and Galeda Bible Class Magazine, September 1917

Let’s practice!



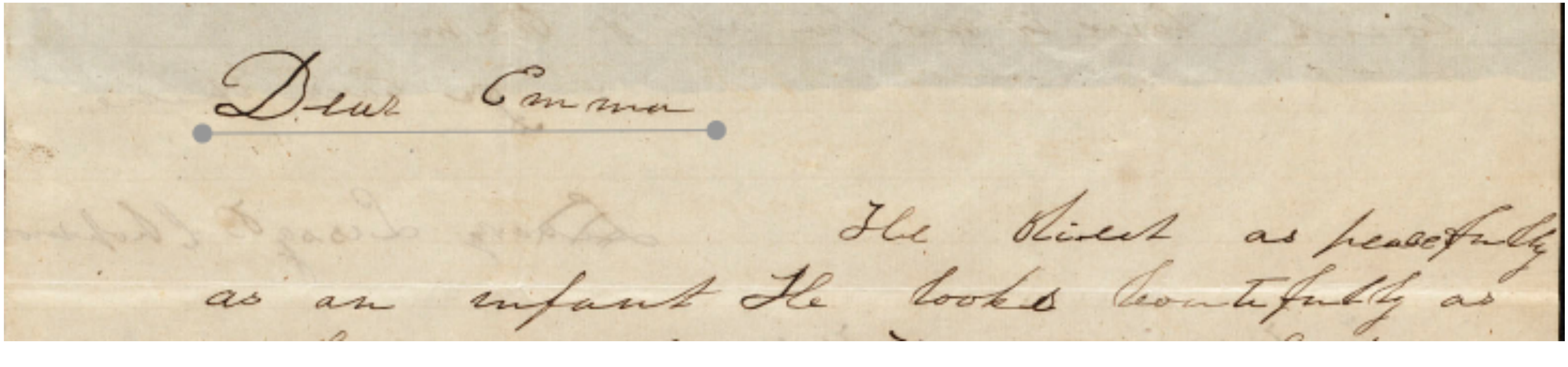
Smithsonian Transcription Center:
Herbarium Project: Santalaceae

Let’s practice!



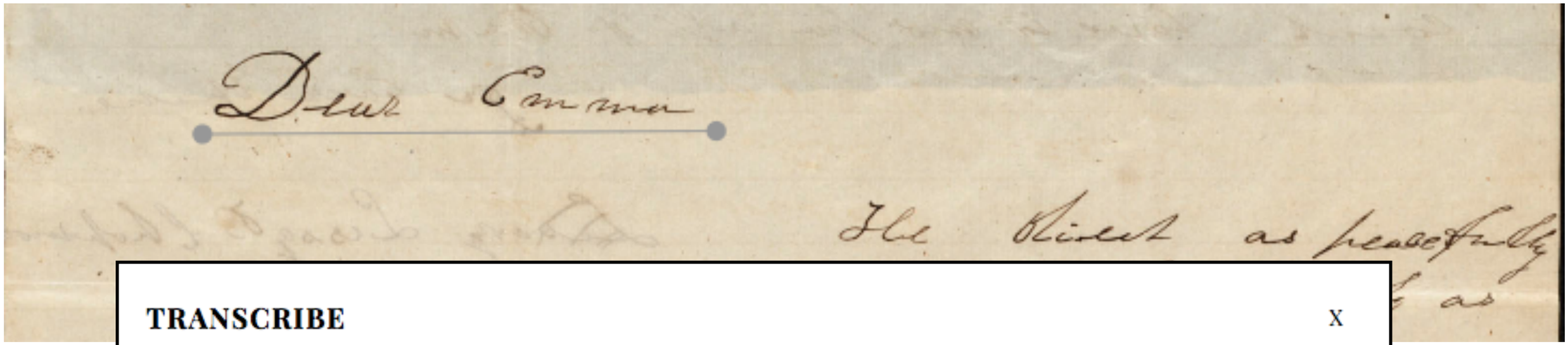
Smithsonian Transcription Center:
Herbarium Project: Santalaceae

Let's practice!



Anti-Slavery Manuscripts
antislaverymanuscripts.org
Zooniverse & Boston Public Library

Let's practice!



TRANSCRIBE X

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

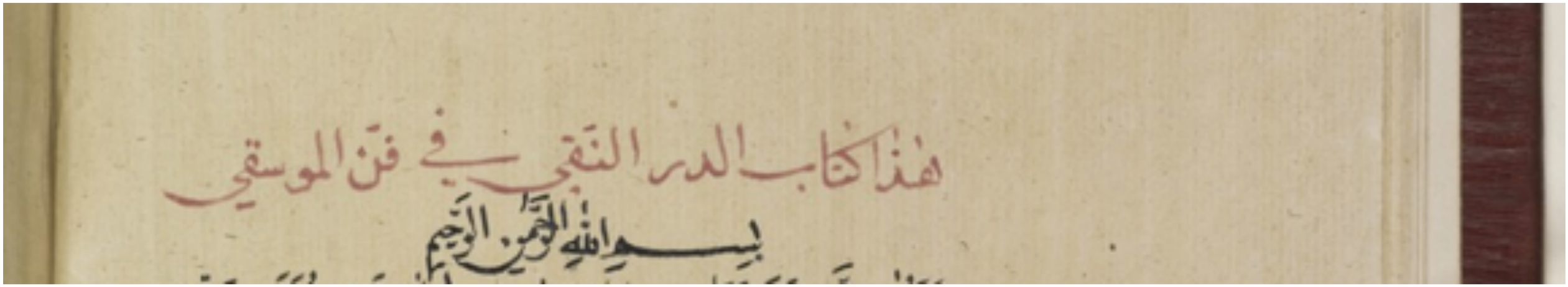
TEXT MODIFIERS [insertion] [deletion] [unclear] [underline]

Dear Emma

Delete Cancel Done

Anti-Slavery Manuscripts
antislaverymanuscripts.org
Zooniverse & Boston Public Library

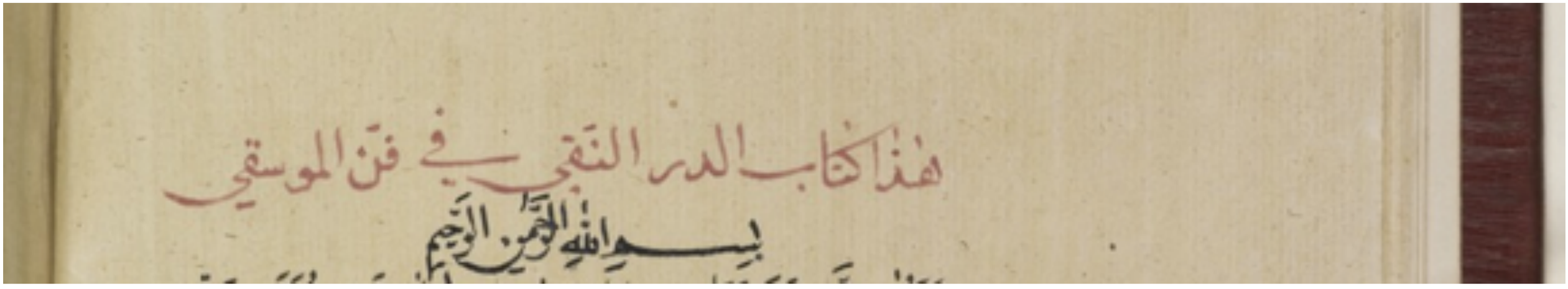
Let’s practice!



al-Durr al-naqī fī fann al-mūsīqī الدرّ النقيّ في الفنّ الموسيقي Mawṣilī, Aḥmad ibn ʿAbd al-Raḥman موصلي، أحمد بن عبد الرحمن [2v] (27/70)

From The Page

Let’s practice!



[1] هذا كتاب الدر النّقي في فنّ الموسقي
[2] بِسم الله الرحمن الرحيم

al-Durr al-naqī fī fann al-mūsīqī الدرّ النقيّ في الفنّ الموسيقي Mawṣilī, Aḥmad ibn ʿAbd al-Raḥman موصلي، أحمد بن عبد الرحمن [2v] (27/70)

From The Page

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

The Codex Mendoza

Tlacuilo: Francisco Gualpuyogualcal

Transcription: Juan González

The Florentine Codex

Misrendered images reflect imperfect literacy

Colonial Transcription

El Título de Santa Maria Ixhuatán (17th c.)

***Teohuanhuaco**:*
Tacuilol pochot tacuilol
teuopixqui
(dibujo de una ceiba, dibujo
de un sacerdote)

Teowakwawnawako
lugar al lado del árbol divino

Teokwawko
lugar del árbol divino

19th Century Transcription

trajeron[3] fueron[4] SS. principales, mas — 3. 4 trajeron – K

... que los gobernaban gobernaron[4] por capitanes. Los de — 4 gobernaban – K

Culiba parecieron[5] gente de mas cuenta y — 5 Asi K. y el MS.

SS. principales. Los unos y los otros vini-

eron á la Laguna de Mexico. Los de Cu- — Laguna de Mexico

hua liba entraron por la parte de Oriente, y

edificaron un pueblo que se dice Tulancin- — Tulancingo Pueblo

Tulancingo) ... leguas de Mexico, y de alli

Tollan (hoy Tula) fueron á Tula, doce leguas de Mexico, á la

parte del Norte, y vinieron poblando hacia

Texcoco Tescuco, que es en la orilla del agua de la La-

guna de Mexico, cinco leguas de traves está

y ocho debajo de de bojeo. Texcoco Tescuco está la parte de

Oriente, y Mexico al Occidente, la laguna

en medio. Algunos quieren decir que Tex Te-

scuco se dice culhua Culiba por respeto de estos

que alli poblaron. Despues el Señorío de

Tescuco) Texcoco Tescuco fué tan grande como el de Mexico.

gobernaban

Culhua

tollantrino

tollan (hoy Tula)

Texcoco

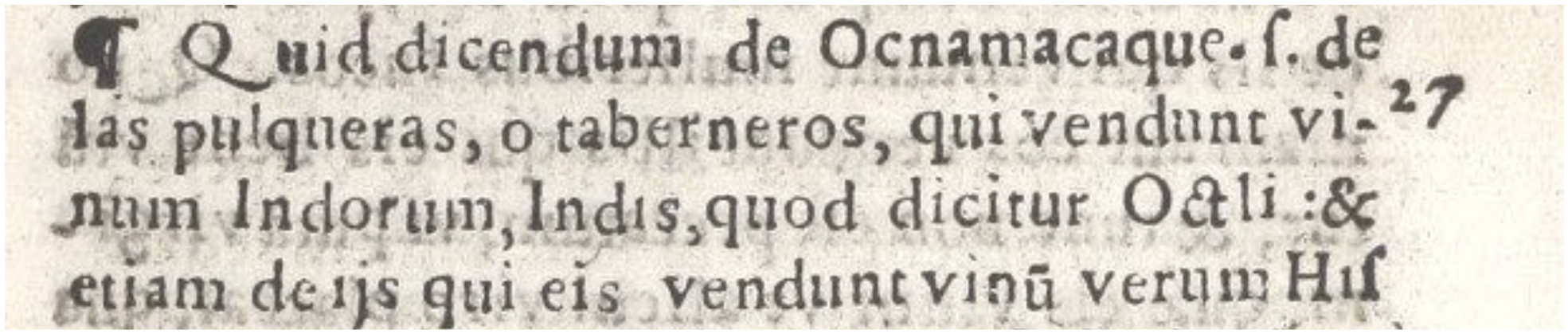
Culhua

Texcoco

Texcoco

21st Century Transcription

"¶ 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 ijs qui eis vendunt vinũ verum Hiſ

Why transcribe?

Accessibility
Discoverability
Preservation
Research

Why transcribe?

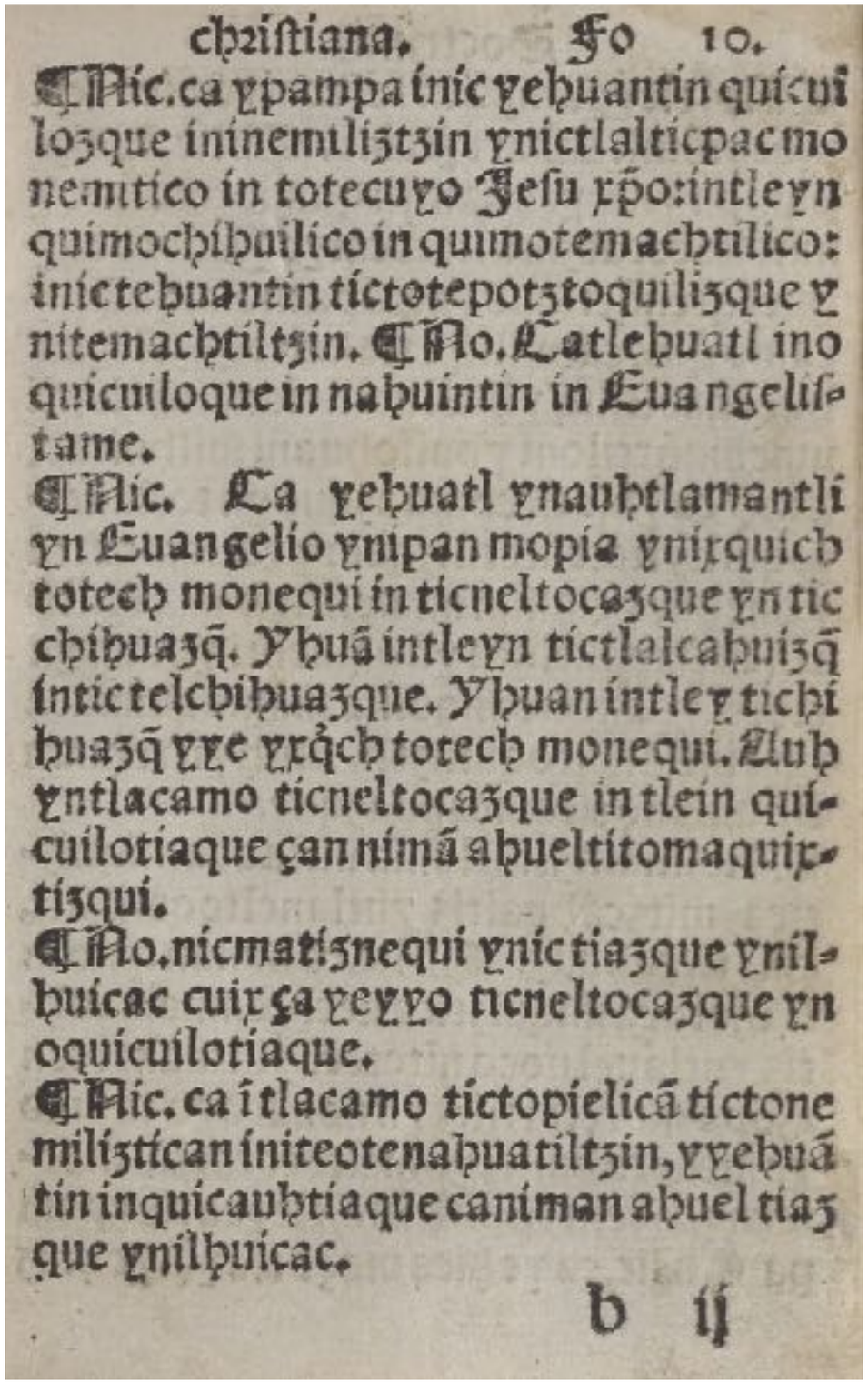
Labor
Cost
Skill
Accuracy
Privacy

What is transcription?

Why transcribe?

Transcribing by machine

Transcribing by hand



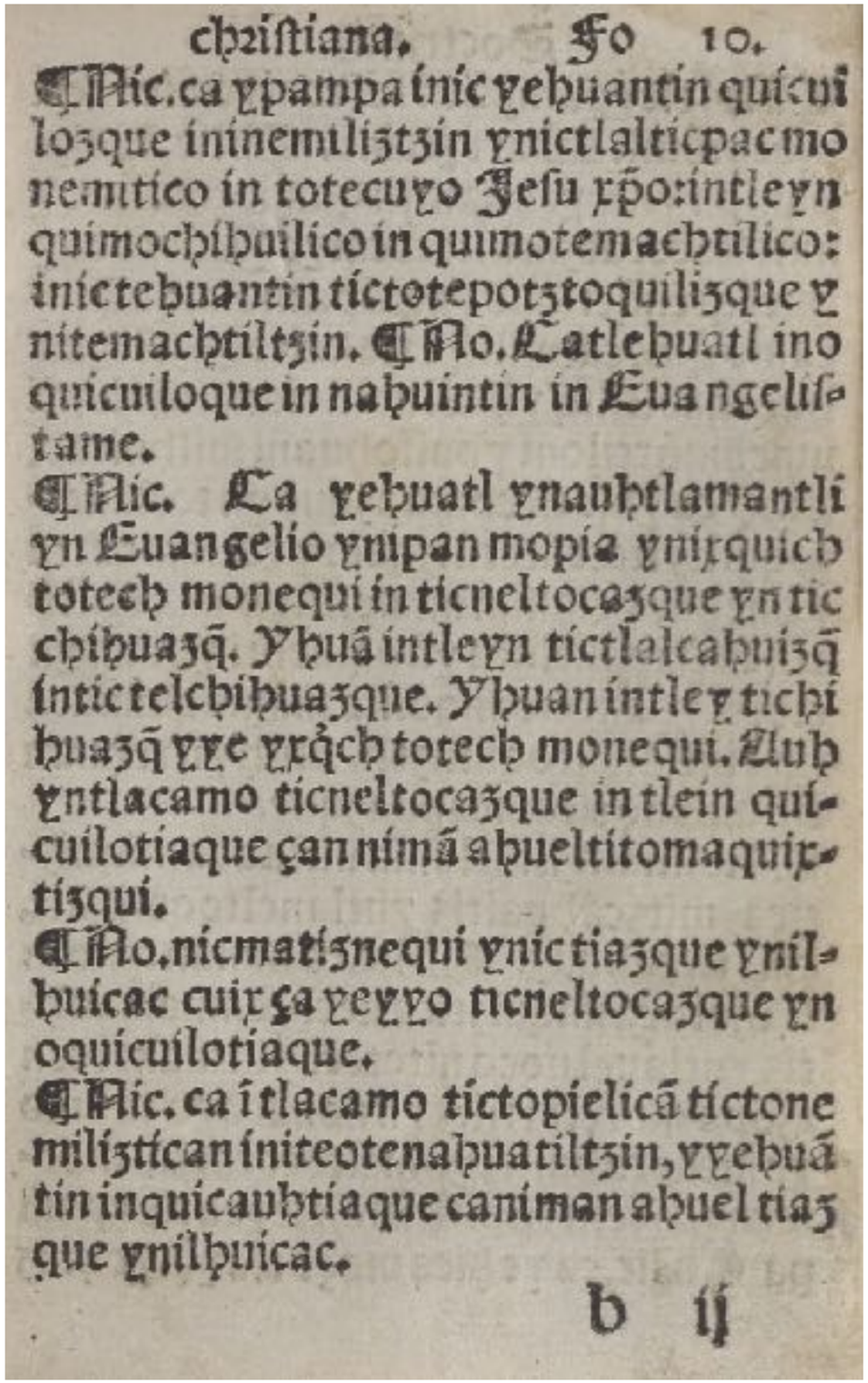
christiana. Fo 10.
¶Nic. ca ypampa inic yehuantin quicui lozque ininemilíztzin ynictlalticpacmo nemitico in totecuyo Jesu xpõ: intleyn quimochihuilico in quimotemachtilico: inictehuantin tictotepotztoquilizque y nitemachtiltzin. ¶No. Catlehuatl ino quicuiloque in nahuintin in Euangelis- tame.
¶Nic. Ca yehuatl ynauhtlamantli yn Euangelio ynipan mopia ynixquich totech monequi in ticneltocazque yn tic chihuazq̃. Yhuã intleyn tictlalcahuizq̃ íntictelchihuazque. Yhuan intley tichi huazq̃ yye yxq̀ch totech monequi. Tluh yntlacamo ticneltocazque in tlein qui- cuilotiaque çan nimã ahueltitomaquix- tizqui.
¶No. nicmatiznequi ynic tiazque ynil- huicac cuix ça yeyyo ticneltocazque yn oquicuilotiaque.
¶Nic. ca îtlacamo tictopielicã tictone miliztican íniteotenahuatiltzin, yyehuã tin inquicauhtiaque caniman ahuel tiaz que ynilhuicac.

b ij

chri|tiana. Fo 10.
¶Nic. ca ypampa inic yehuantin quicui lozque ininemilztzin yn ictlalticpacmo nemitico in totecuyo Je|u x\~po: intleyn quimochihuilico in quimotemachtilico: inictehuantin tictotepotztoquilizque y nitemachtiltzin. ¶No. Catlehuatl ino quicuiloque in nahuintin in Euangeli|- tame.
¶Nic. Ca yehuatl yn auhtlamantli yn Euangelio yn ipan mopia yn ixquich totech monequi in ticneltocazque yn tic chihuaz\~q. Yhu\~a intleyn tictlalcahuiz\~q íntictelchihuazque. Yhuan intley tichi huaz\~q yye yx\'qch totech monequi. Tluh yntlacamo ticneltocazque in tlein qui- cuilotiaque çan nim\~a ahueltitomaquix- tizqui.
¶No. nicmatiznequi yn ic tiazque yn il- huicac cuix ça yeyyo ticneltocazque yn oquicuilotiaque.
¶Nic. ca \~itlacamo tictopielic\~a tictone miliztican initeotenahuatiltzin, yyehu\~a tin inquicauhtiaque caniman ahuel tiaz que yn ilhuicac.

b ij

Optical Character Recognition | OCR


christiana. Fo 10.
¶ Nic.ca ypampa inic yehuantin quicui
lozque ininemiliztzin ynictlalticpac mo
nemitico in totecuyo Jesu xpo:intleyn
quimochihuilico in quimotemachtilico:
inictehuantin tictotepotztoquilizque y
nitemachtiltzin. ¶ No. Catlehuatl ino
quicuiloque in nahuintin in Euangelis-
tame.
¶ Nic. Ca yehuatl ynauhtlamantli
yn Euangelio ynipan mopia ynixquich
totech monequi in ticneltocazque yn tic
chihuazq̃. Yhuã intleyn tictlalcahuizq̃
inticteIchihuazque. Yhuan intley tichi
huazq̃ yye yrq̀ch totech monequi. Auh
yntlacamo ticneltocazque in tlein qui-
cuilotiaque çan nimã ahueltitomaquix-
tizqui.
¶ No. nicmatiznequi ynictiazque ynil-
huicac cuix ça yeyyo ticneltocazque yn
oquicuilotiaque.
¶ Nic. ca itlacamo tictopielicã tictone
milizticaniniteotenahuatiltzin, yyehuã
tin inquicauhtiaque caniman ahueltiaz
que ynilhuicac.
b ij

o — M r^ -^ v^SO |>»00 Cv O — «
r^ -«t- «^\o 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^j^j^c^r>.c^r>. r**oo

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

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

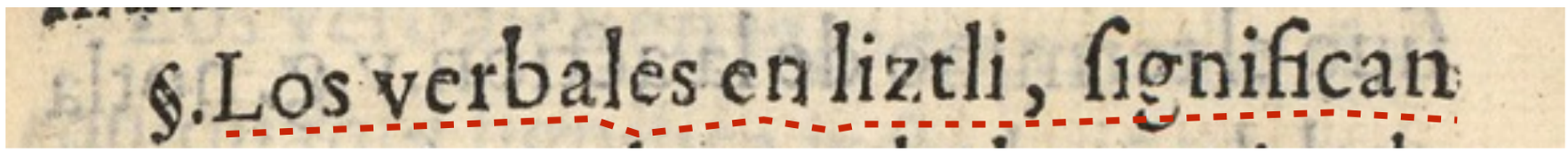
00 (7s O — «*< '♦N ^ w^sO r-»00

(7s o «- n r^ 't-

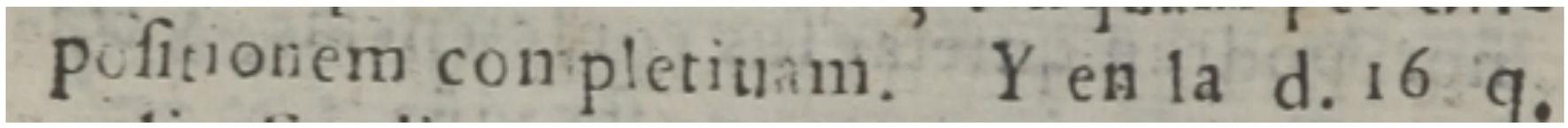
Optical Character Recognition | OCR

Material Challenges

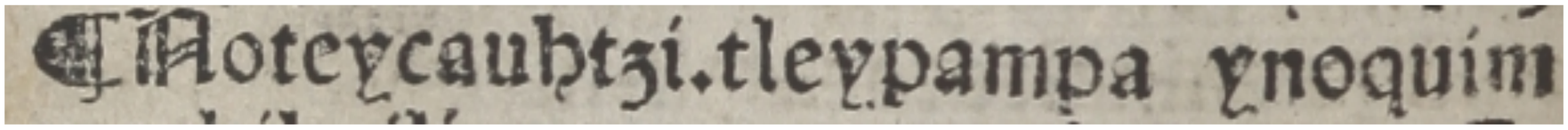
Wandering baseline



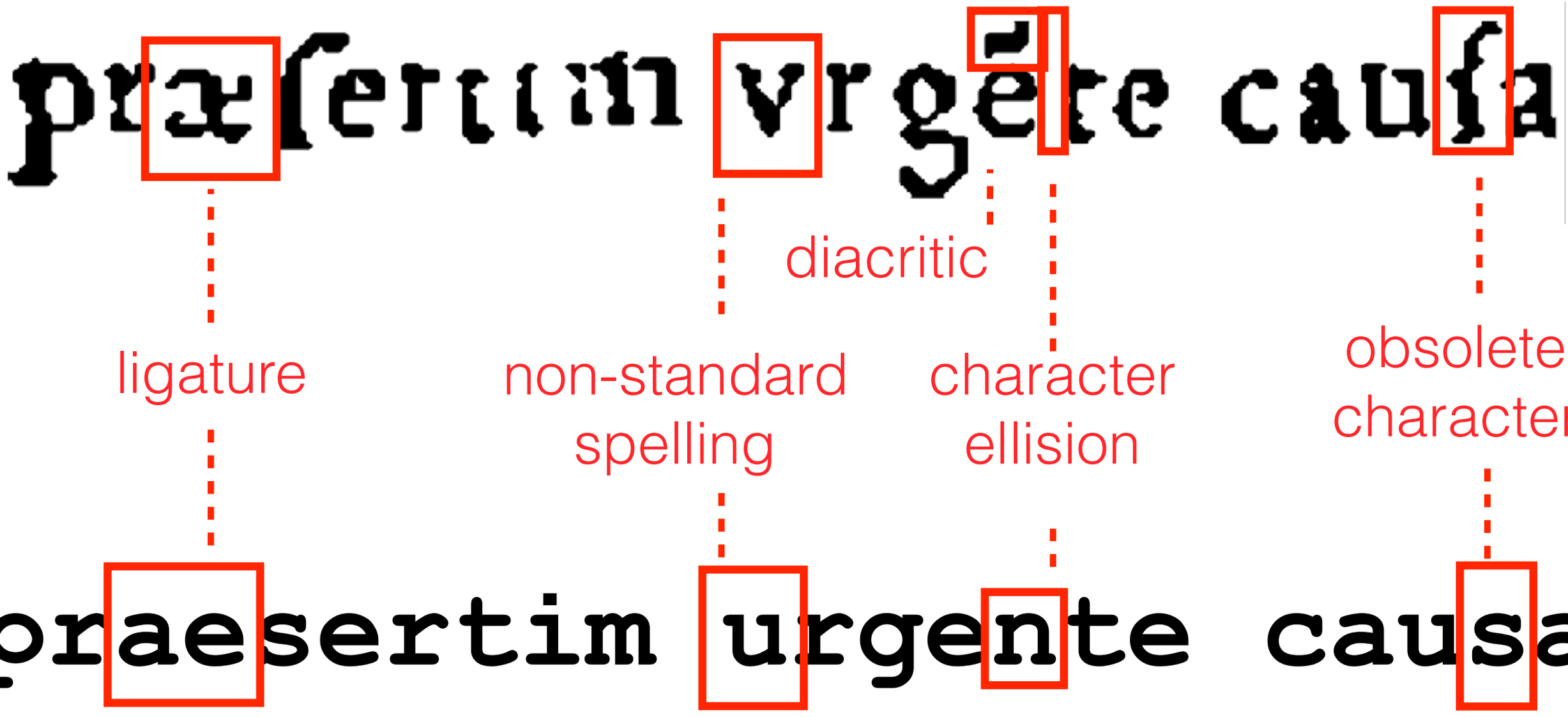
Uneven inking



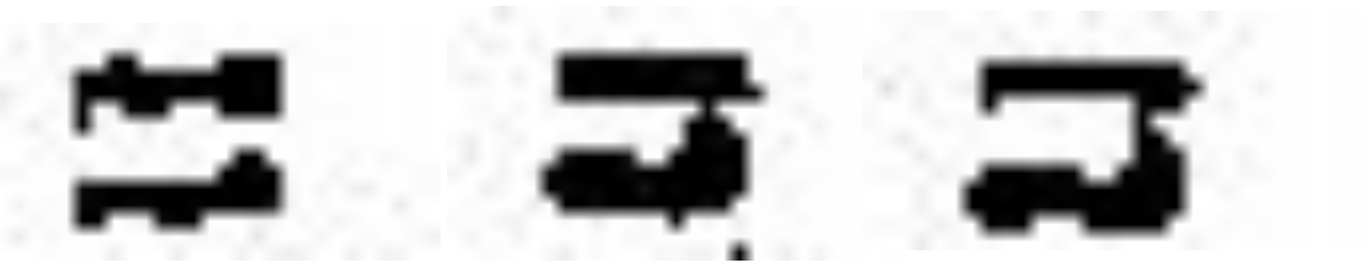
Unfamiliar Typefaces

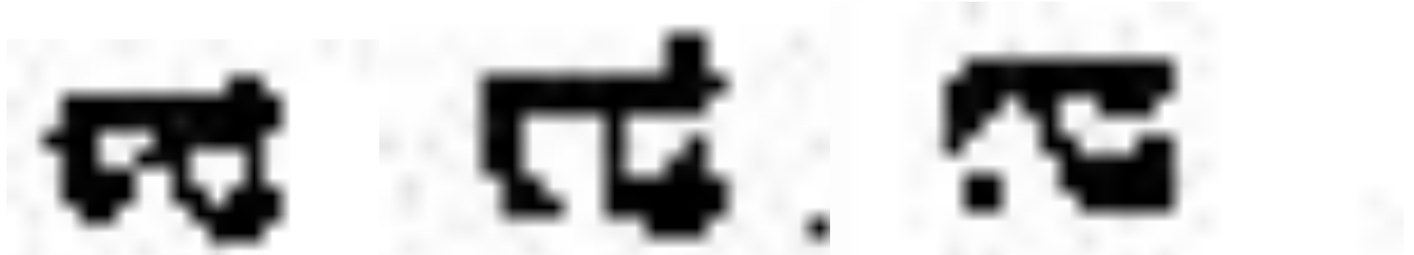


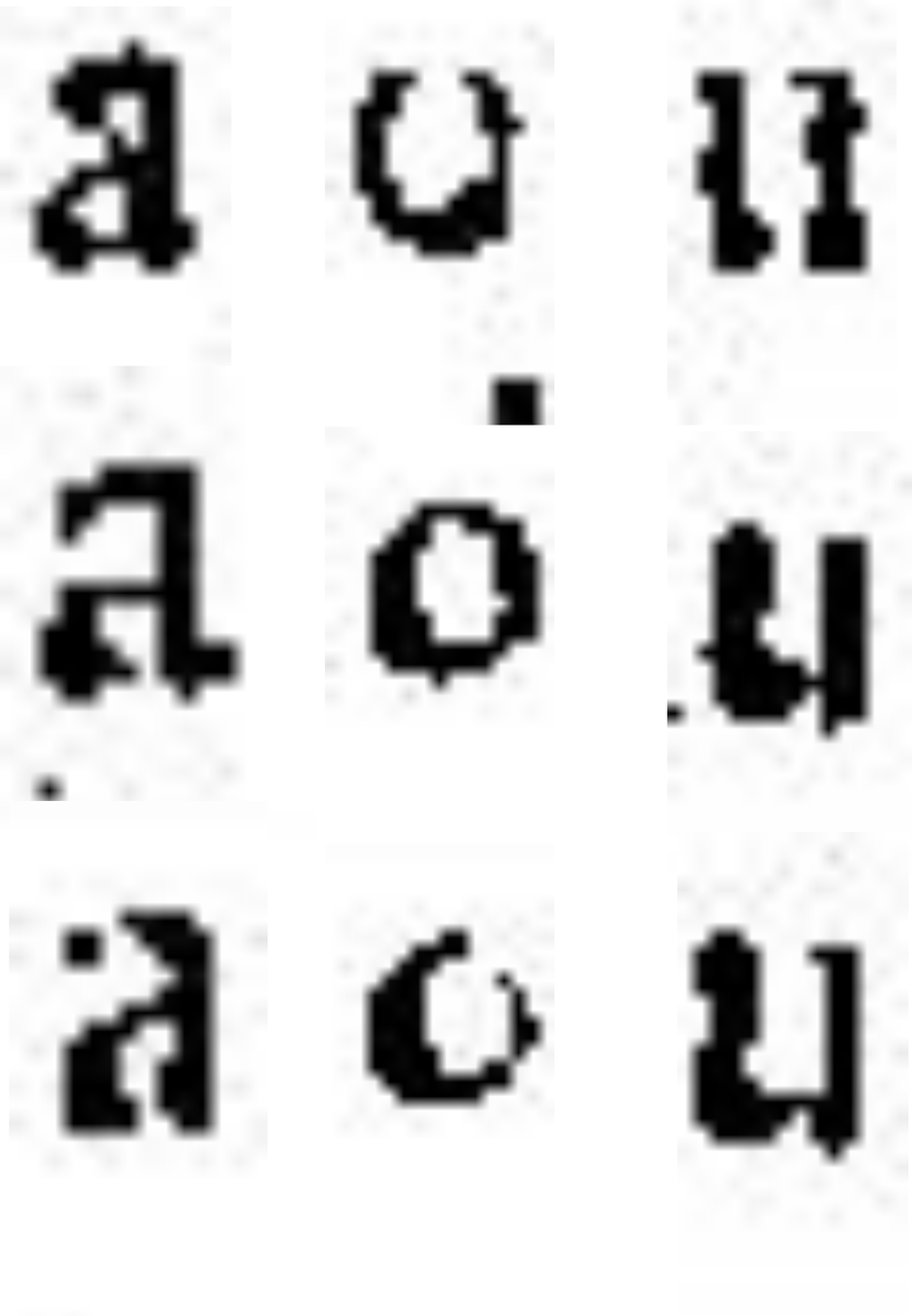
Orthographic Challenges



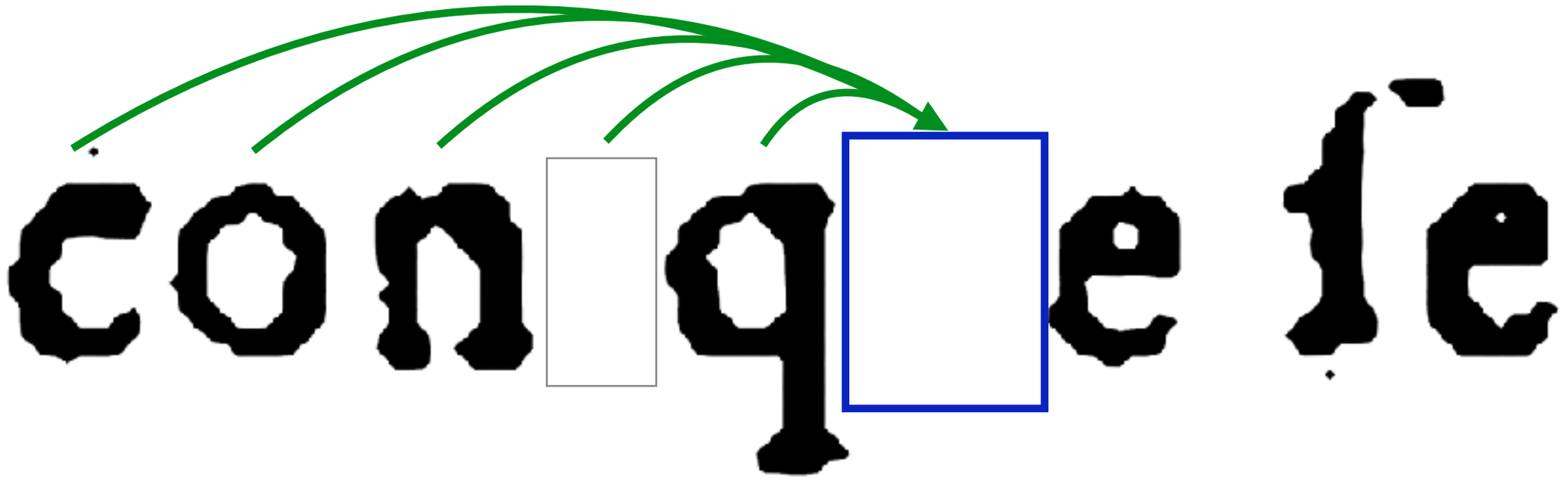
How does OCR work

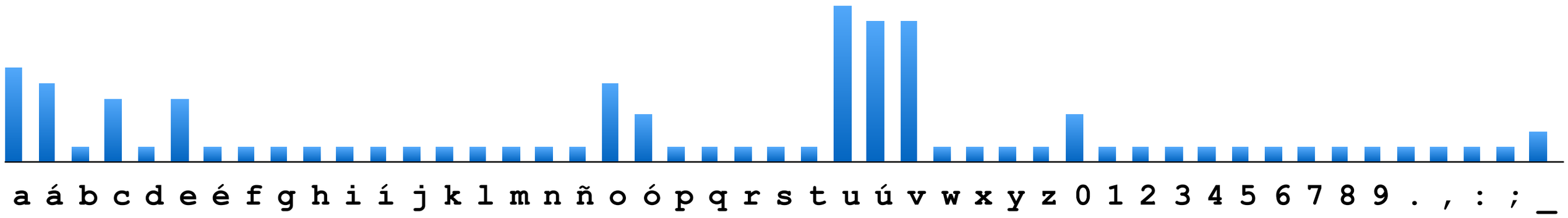






a á b c d e é f g h i í j k l m n ñ o ó p q r s t u ú v w x y z 0 1 2 3 4 5 6 7 8 9 . , : ;





a á b c d e é f g h i 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 fe

con que fe

Options for Automatic Transcription

Et at id quod obiicitur Clementem Septimũ ſolum fuiſſe locutum de priuilegiis cõceden dis ordinibus ſuo tempore a Sede Apoſtoli- ca approbatis, & religionem SOCIETA TIS IESV multo poſt, fuiſſe a Sede Apo- ſtolica approbatam : reſponder, quòd poſt Clementem Septimum, Paulus Quartus, Pi us Quartus, & Pius Quintus, & Gregori- us Decimus tertius, & Sixtus Quintus priui legia Mendicantium approbantes, ea õnia de nouo concefferunt : ac ſi tenor ipſorum priui legiorum in eorum literis Apoſtolicis appo- neretur, & ſic priuilegium Clementis Septi- mi tempore & ætate horum Summorum Pon tificum fuiſſe conceſſum, quo tempore, & æta te, religio SOCIETATIS IESV nõ ſolum erat ab Eccleſia approbata, ſed etiam per orbem diſperſa. Verum hoc eſt certum & indubitatum Adrianum Sextum Pontificẽ maximum ad inſtantiam Caroli maximi Ro- manorum Imperatoris & Hiſpaniarum Regis conceſsiſſe fratribus minoribus de obſeruan tia exiſtentibus, aut ire procurantibus, in di-

114

Los Con feffores, A Er atid quod obiicitur Clementem Septinit folum fuiffe locutum de priuilegiis coceden dis ordinibus fuo tempore a Sede Apoftoli ca approbatis, & religionem SOCIETA TISHTE S Va multo poft, fuiffe a Sede Apo ftolica approbatam : respondet quòd poft Clementem Septimum, Paulus Q uartus, PL us Q uartus, & Pius Q uintus, & Gregori tis Decimtis tertius, & Sixtus Quintus priui legia Mendicantium approbantes, ea õnia de Holo concefferunt : ac fi tenor ipforum priui legiorum in eorum literis Apoftolicis appo neretur, & fic priuilegium Clementis Septi. mi tempore & ætate horiim Summorum Pon trficum fuisse conceffum, quo tempore, &c æ ta te, religio S O CIETATIS IESV. nó folyui erat ab Ecclefia approbata, fed etiam Per orbem dispersa. Verum hoc est certum & indubitatum Adrianlin Sexilm Pontifice maximum ad inftantiam Caroli maximi Ro- manorum Imperatoris & Hispaniarum Regis

concessiffe fratribus minoribus de obseruan tia exiftentibus, aut ire procurantibus, in die

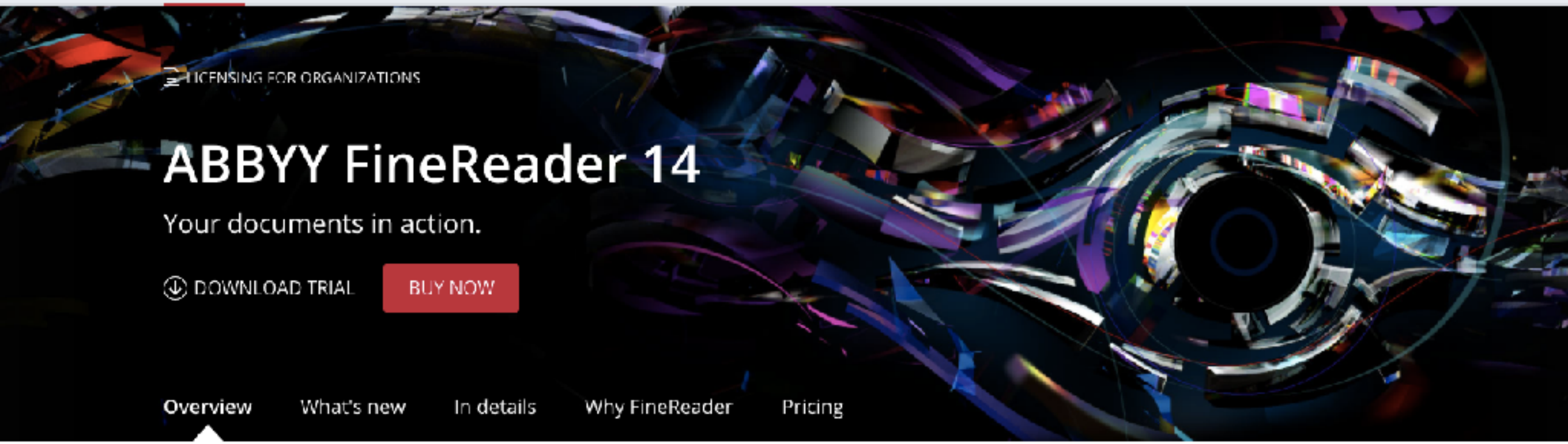
Free
Easy
Limited Quantity
No training or modification

Google Drive

ABBYY

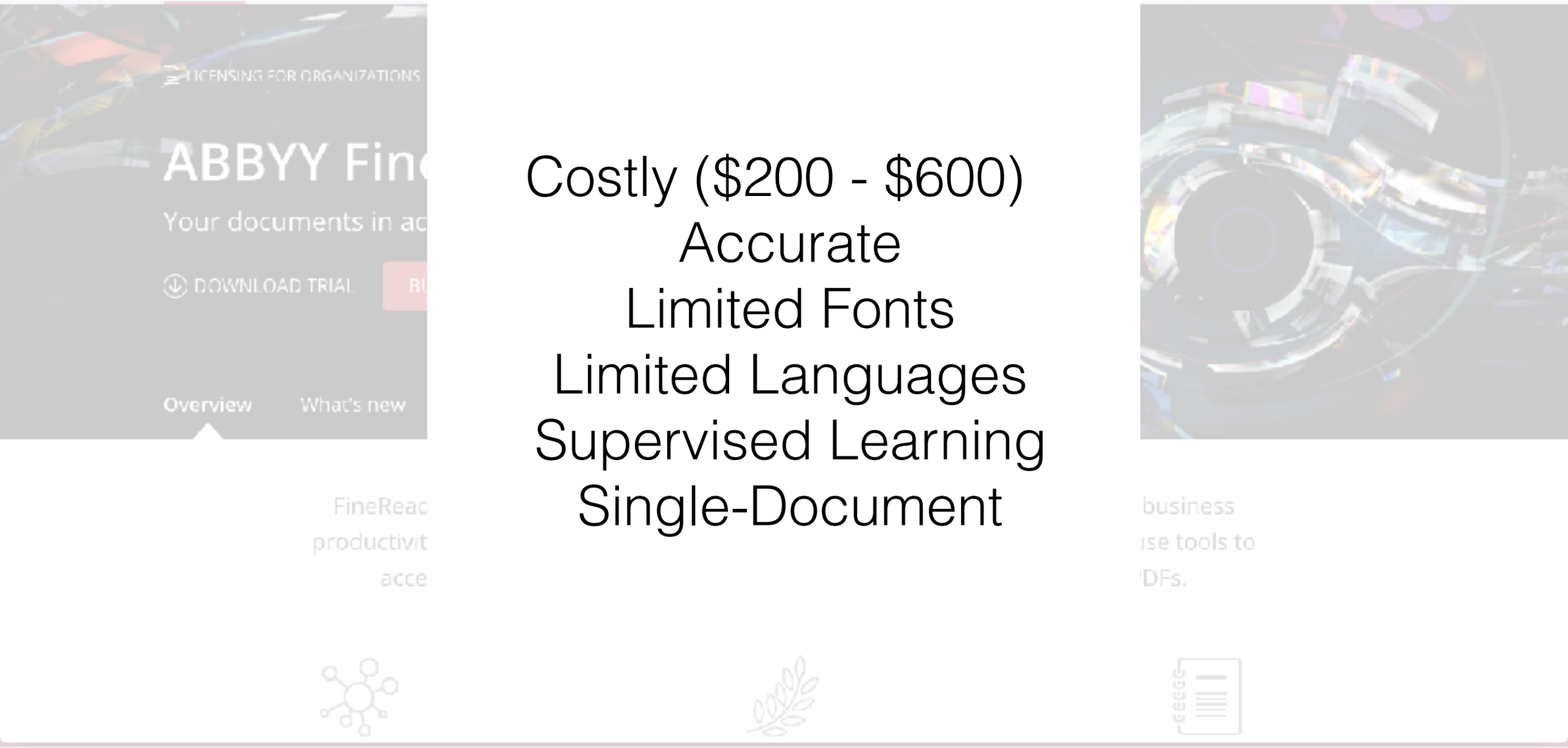
CONTACTS STORE SELECT REGION

Products Solutions Developer Tools Partners Support Company



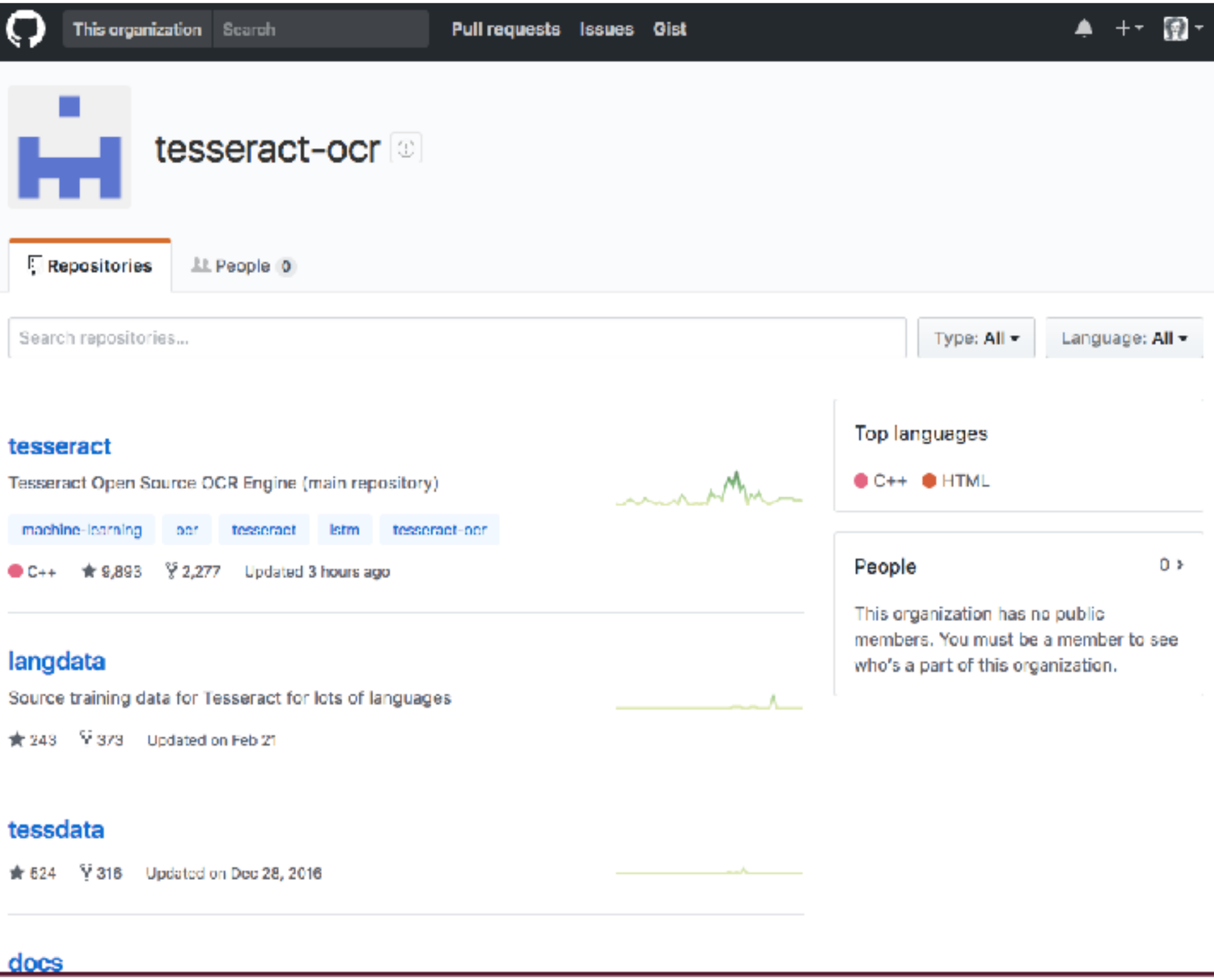
FineReader is an all-in-one OCR and PDF software application for increasing business productivity when working with documents. It provides powerful, yet easy-to-use tools to access and modify information locked in paper-based documents and PDFs.

ABBYY Finereader



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

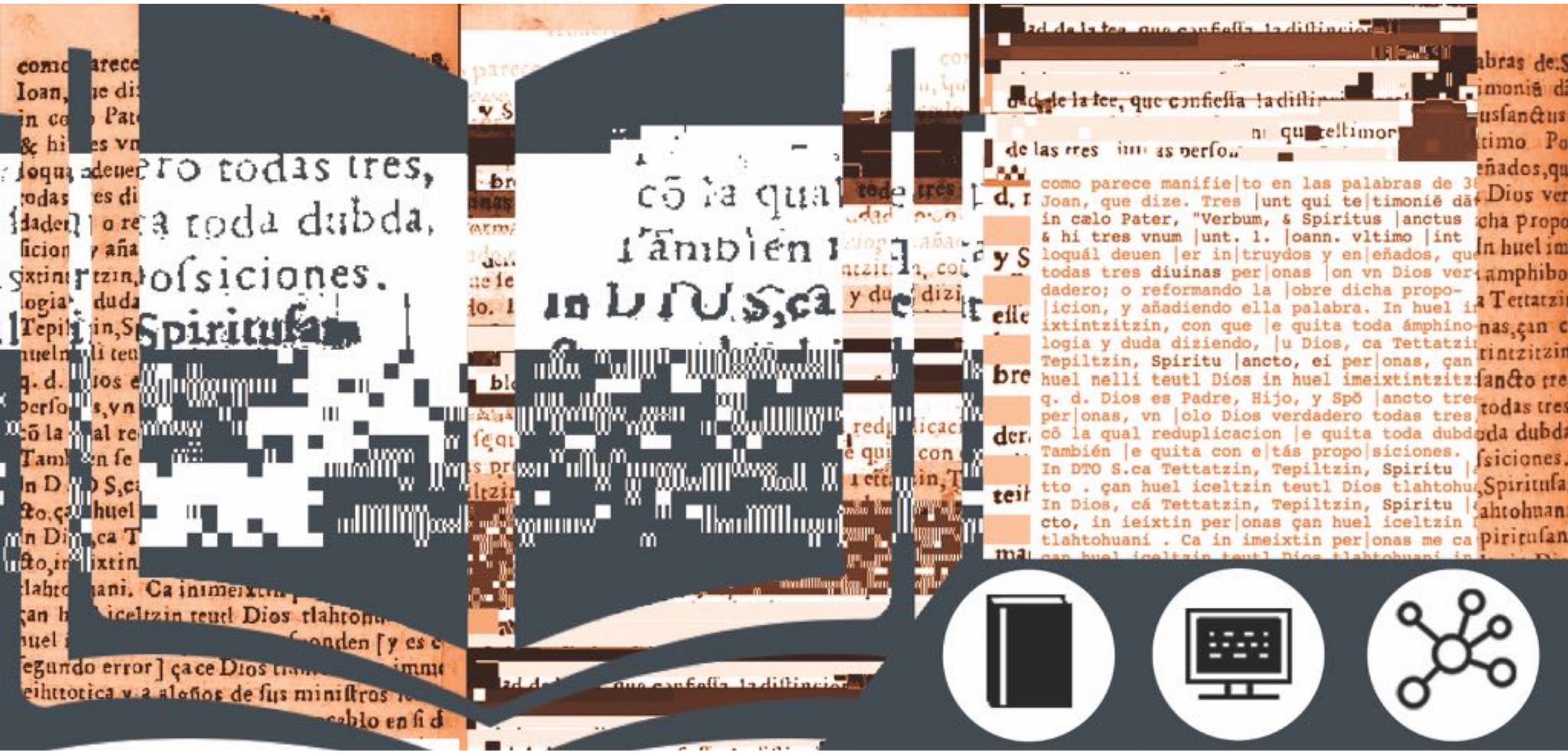
ABBYY Finereader


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
langdata
Source training data for Tesseract for lots of languages
243 373 Updated on Feb 27
tessdata
624 316 Updated on Dec 28, 2016
docs
Top languages
C++ HTML
People 0 ›
This organization has no public members. You must be a member to see who's a part of this organization.

Tesseract

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

Tesseract


como parece manifie|to en las palabras de 3
Joan, que dize. Tres |unt qui te|timonié dā
in cælo Pater, "Verbum, & Spiritus |anctus
& hi tres vnum |unt. 1. |oann. vltimo |int
loquál deuen |er in|truydos y en|eñados, qu
todas tres diuinas per|onas |on vn Dios ver
dadero; o reformando la |obre dicha propo-
|icion, y añadiendo ella palabra. In huel i
ixtintzitzin, con que |e quita toda ámphino-
logia y duda diziendo, |u Dios, ca Tettatzi
Tepiltzin, Spiritu |ancto, ei per|onas, çan
huel nelli teutl Dios in huel imeixtintzitz
q. d. Dios es Padre, Hijo, y Spō |ancto tre
per|onas, vn |olo Dios verdadero todas tres
cō la qual reduplicacion |e quita toda dubd
También |e quita con e|tás propo|siciones.
In DTO S.ca Tettatzin, Tepiltzin, Spiritu |
tto . çan huel iceltzin teutl Dios tlahtohu
In Dios, cá Tettatzin, Tepiltzin, Spiritu |
cto, in ieixtin per|onas çan huel iceltzin
tlahtohuani . Ca in imeixtin per|onas me ca

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
primeroslibros.org
emop.tamu.edu

The Primeros Libros Project

21 Partner Institutions
430 Exemplars
8 Languages

www.primeroslibros.org

Reading the First Books Project Goals:

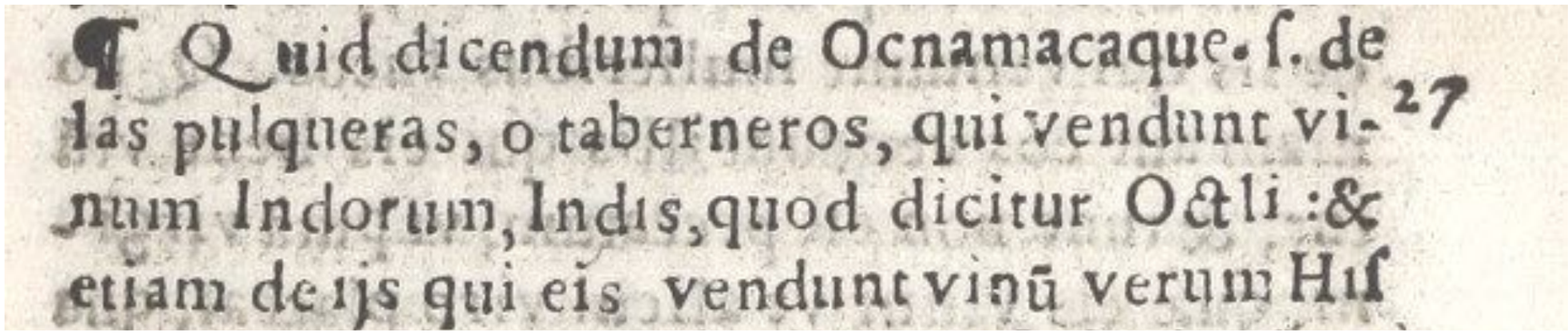
1. **Develop:** transcription tools for multilingual early-modern texts

2. **Integrate:** tools with the Early Modern OCR Project interface.

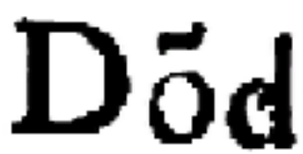
3. **Implement:** transcribe the books in the Primeros Libros project.

Multilingual OCR

"¶ 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

Orthographic Variability

	Original form	Modern form
	di**z**e	di**c**e
	n**u**mero	n**ú**mero
	D**õ**de	D**on**de

Orthographic Variability

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

Orthographic Variability

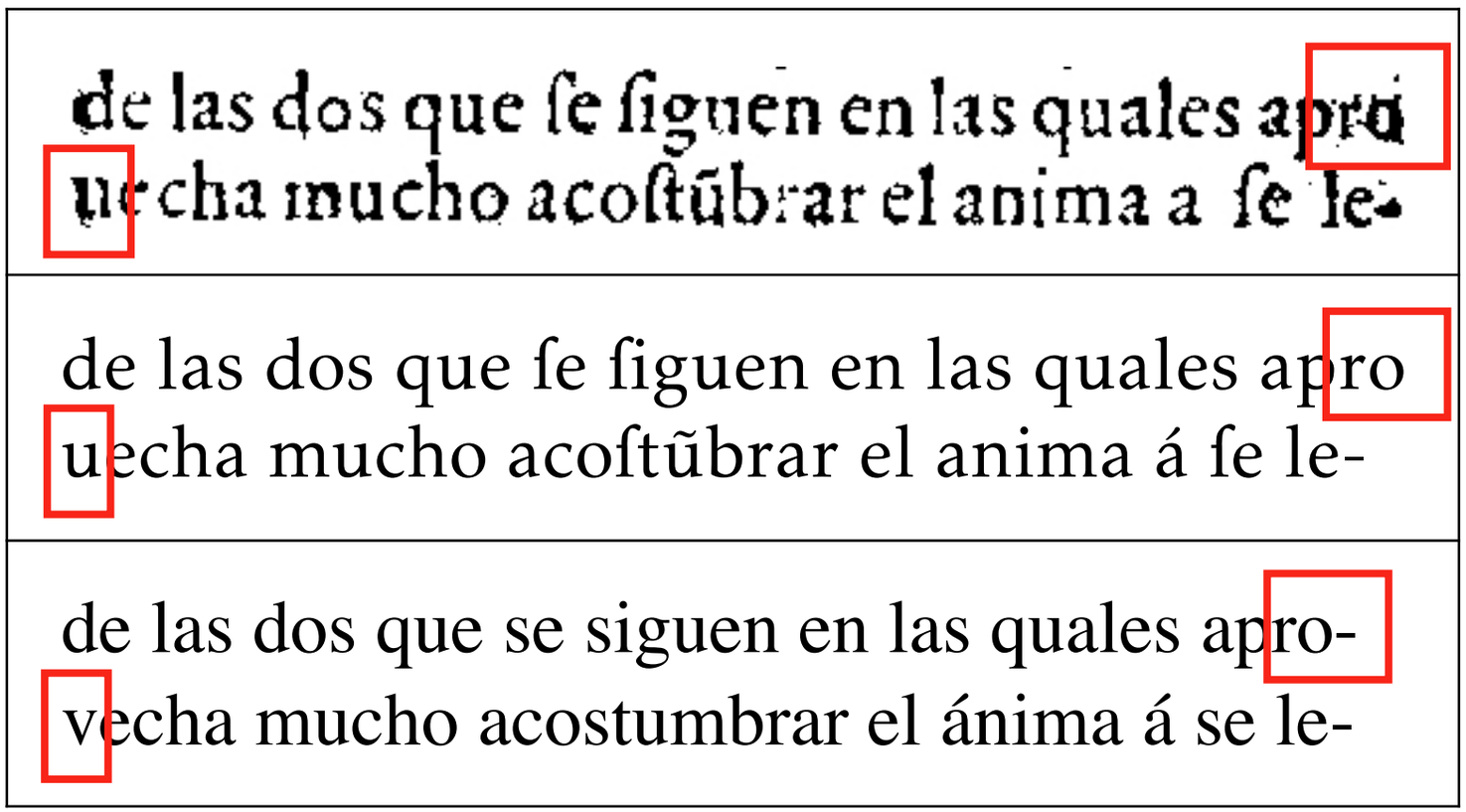
Without handling orth. variation: `merita`

Correct diplomatic transcription: `mẽtira`

Correct normalized form: `mentira`

Orthographic Variability

Original image


de las dos que ſe ſiguen en las quales apro
uecha mucho acoſtũbrar el anima a ſe le-

Our diplomatic

de las dos que ſe ſiguen en las quales apro
uecha mucho acoſtũbrar el anima á ſe le-

Our normalized

de las dos que se siguen en las quales apro-
vecha mucho acostumbrar el ánima á se le-

Orthographic Variability

Original image	
Baseline	Nefuxpo
Our diplomatic	*Jeſuxpo
Our normalized	*Jesuxpo
Gold diplomatic	Jeſu xp̃o
Gold Normalized	Jesu Cristo

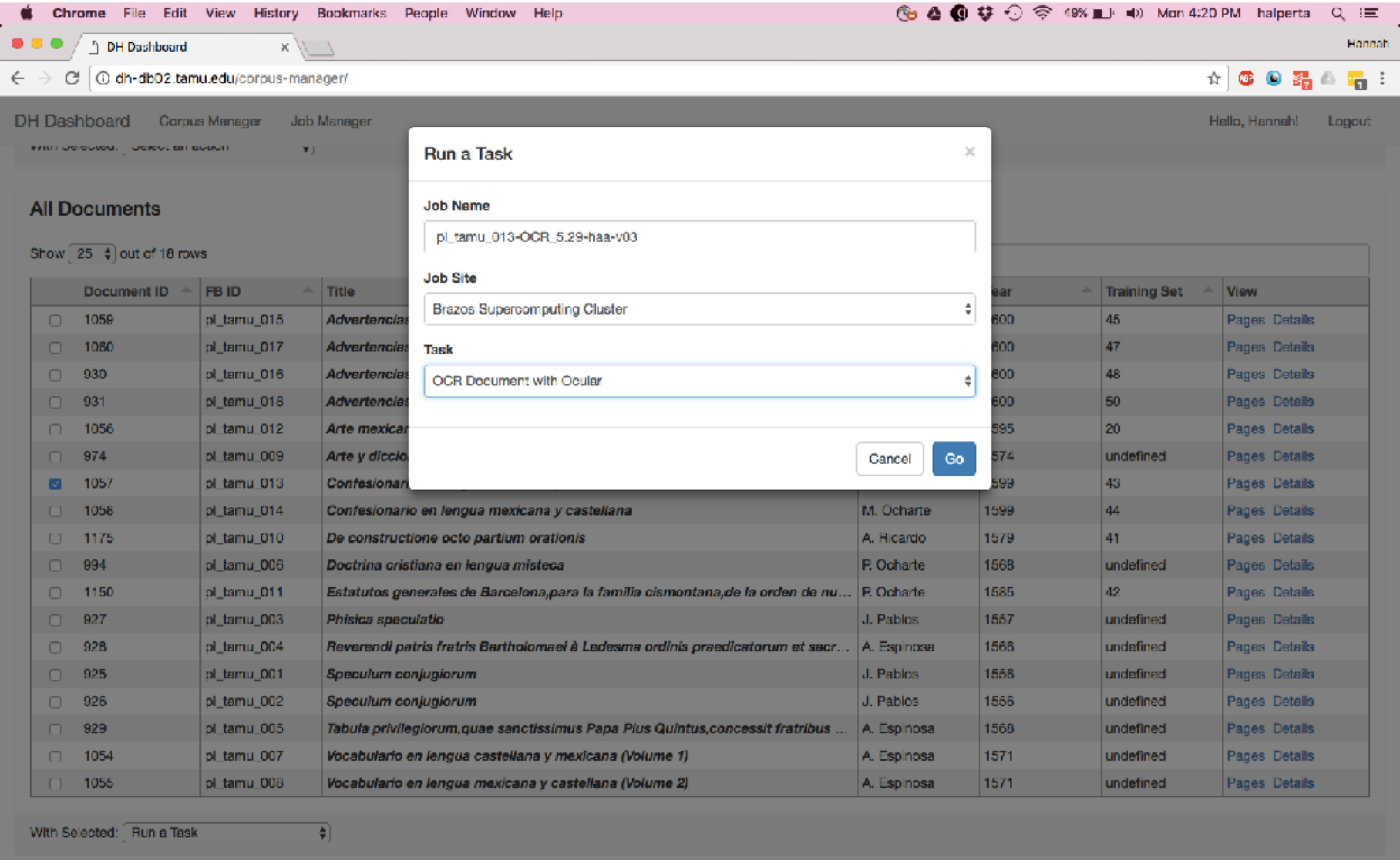
Reading the First Books Project Goals:

1. **Develop:** transcription tools for multilingual early-modern texts

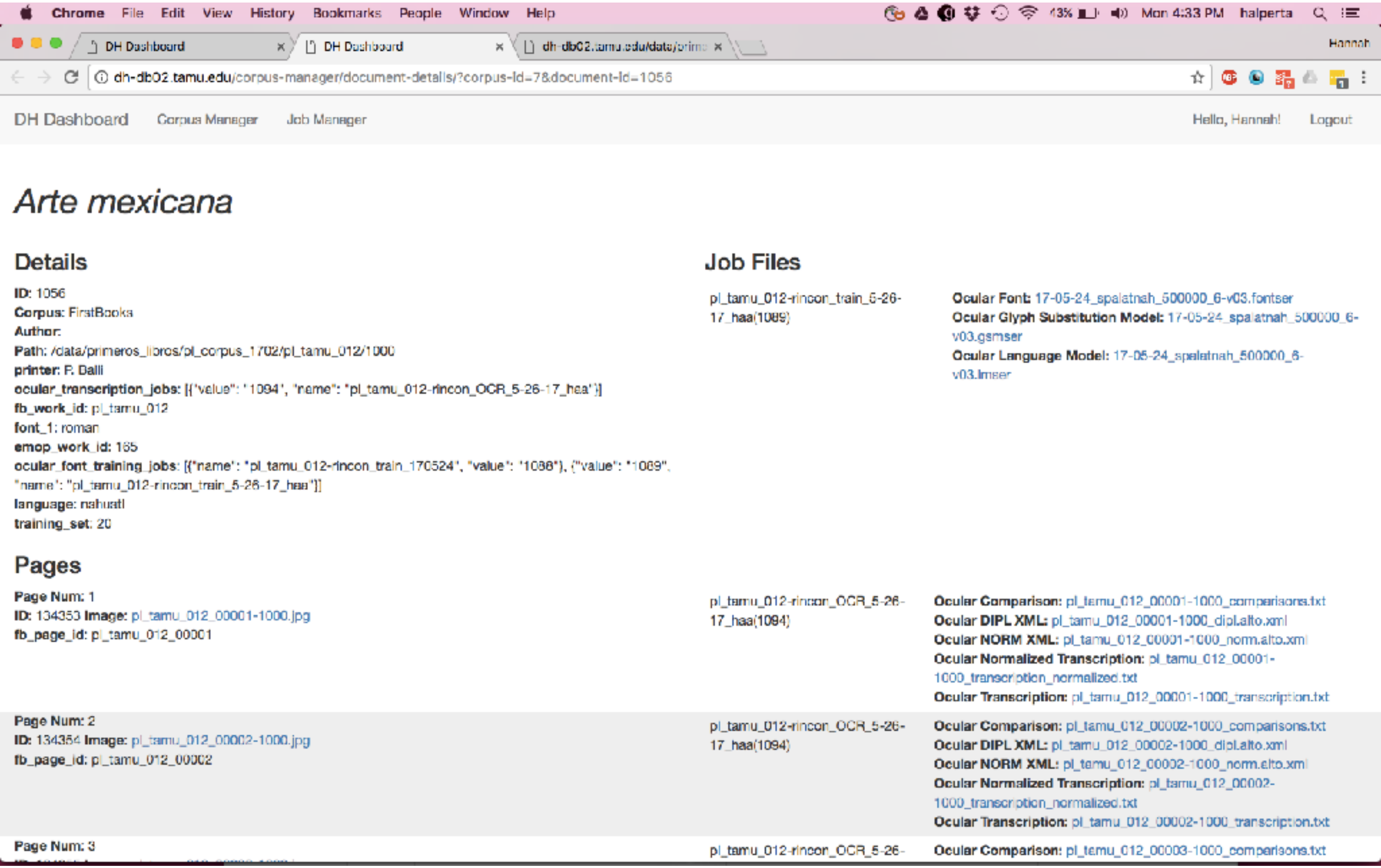
2. **Integrate:** tools with the Early Modern OCR Project interface.

3. **Implement:** transcribe the books in the Primeros Libros project.

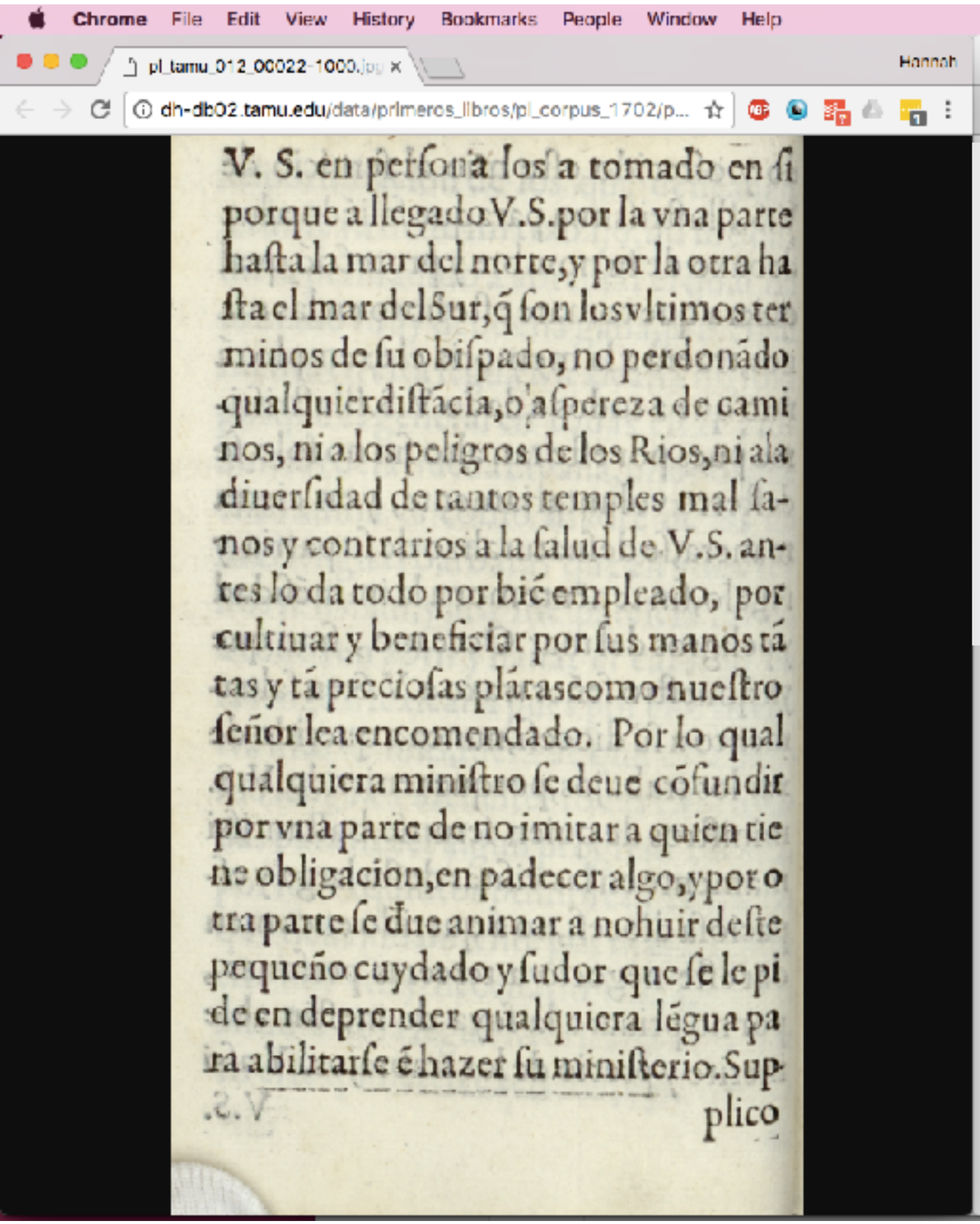
Ocular Interface for eMOP



Ocular Interface for eMOP

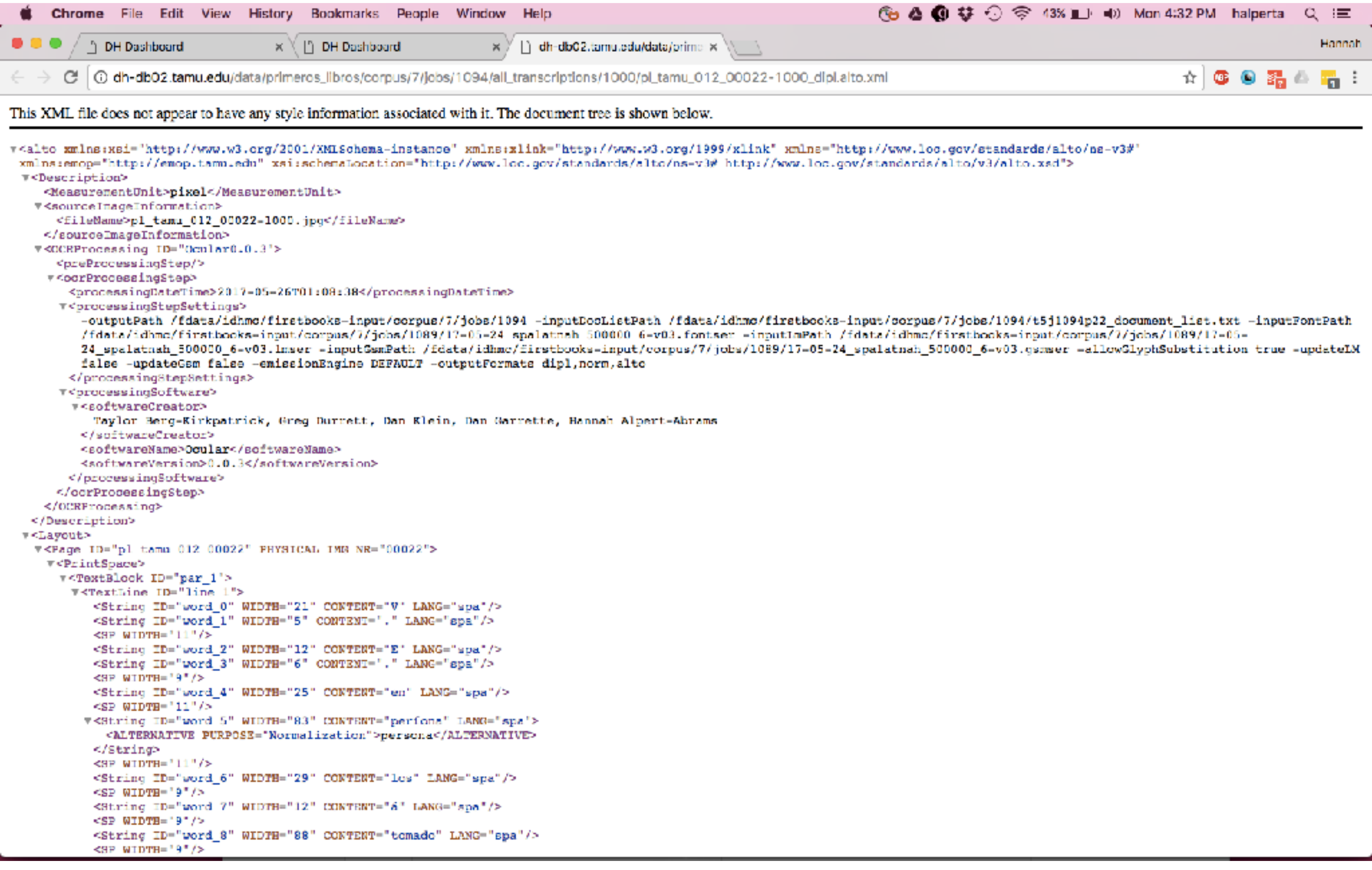


Ocular Interface for eMOP


V. S. en perſona los a 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, q̃ ſon los vltimos ter
minos de ſu obiſpado, no perdonádo
qualquier diſtácia, o aſpereza de cami
nos, ni a los peligros de los Rios, ni ala
diuerſidad de tantos temples mal ſa-
nos y contrarios a la ſalud de V.S. an-
tes lo da todo por bié empleado, por
cultiuar y beneficiar por ſus manos tá
tas y tá precioſas plátas como nueſtro
ſeñor lea encomendado. Por lo qual
qualquiera miniſtro ſe deue cõfundir
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 deſte
pequeño cuydado y ſudor que ſe le pi
de en deprender qualquiera lẽgua pa
ra abilitarſe é hazer ſu miniſterio. Sup-
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 biẽemple ado, por
cultiuar y beneficiar por ſus manos tã
tas y ta precioſas platas como nueſtro
Ieñor lea encomendado. Por lo qual
qualquiera miniſtro ſe deue cófundir
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 cuydado y ſudor que fe le pĩ
de en de prender qualquieta legua Pa
ra abilitarſe ẽ hazer firminiſterio. Su p-
plicó

Ocular Interface for eMOP



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.

Transcribed PL Corpus

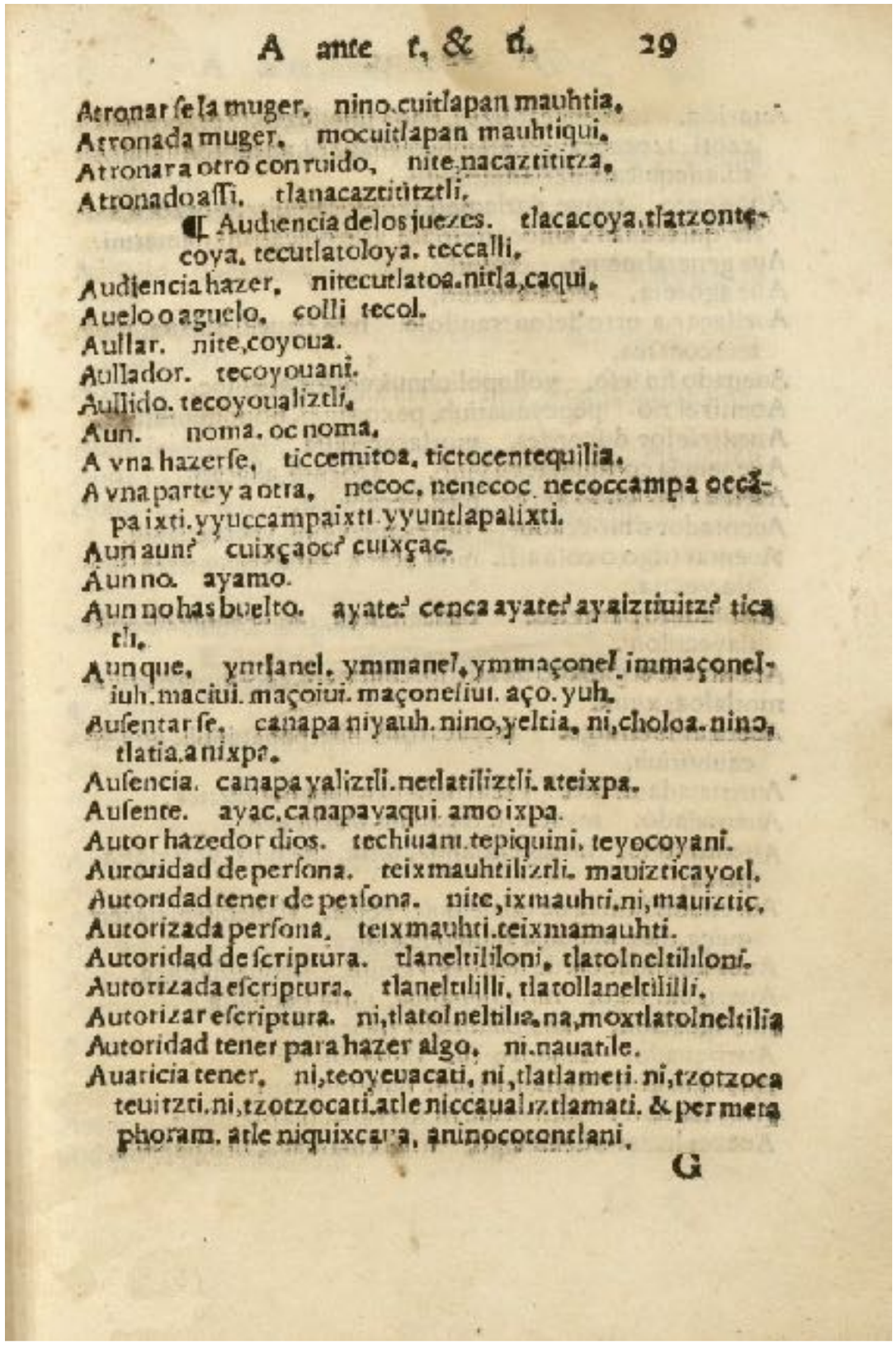
¶ Aqui comiença vn vocabula-
rio en la lengua Caſtellana y Mexicana, Compueſto
por el muy reuerendo padre fray Alonſo de
Molina: Guardiã del conuẽto de ſant Antonio de
Tezcuco de la ordẽ de los frayles Menores.

¶ Idorum nimia te fecit prole parentem.
qui genuit moriens, quos pater alue foues,
Confixus viuis, langues; cum mente reuolues.
vulnera, cum ſpectas, ſtigmata carne geris.

Atronar ſe la muger. nino.cuitlapan mauhtia.
Atronada muger. mocuitlapan mauhtiqui.
Atronar a otro con ruido. nite.nacaztititza.
Atronado aſſi. tlanacaztititztli.
¶ Audiencia de los juezes. tlacacoya.tlatzonte-
coya. tecutlatoloya. teccalli.
Audiencia hazer. nitecutlatoa.nitla.caqui.
Auelo o aguelo. colli tecol.
Aullar. nite.coyoua.
Aullador. tecoyouani.
Aullido. tecoyoualiztli.
Aun. noma. oc noma.
A vna hazerſe. ticcemitoa. tictocentequilia.
A vna parte y a otra. necoc. nenecoc. necoccampa occã-
pa ixti. yyuccampaixti. yyuntlapalixti.
Aun aun? cuixçaoc? cuixçac.
Aun no. ayamo.
Aun no has buelto. ayate? cenca ayate? ayaiztiuitz? tica
tli.
Aunque. yntlanel. ymmanel. ymmaçonel. immaçonel-
iuh. maciui. maçoiui. maçoneſiui. aço. yuh.
Auſentar ſe. canapa niyauh. nino.yeltia. ni.choloa. nino.
tlatia. anixpa.
Auſencia. canapa yaliztli. netlatiliztli. ateixpa.
Auſente. ayac. canapayaqui. amoixpa.
Autor hazedor dios. techiuani. tepiquini. teyocoyani.
Autoridad de perſona. teixmauhtilizrli. mauizticayotl.
Autoridad tener de perſona. nite.ixmauhti.ni.mauiztic.
Autorizada perſona. teixmauhti. teixmamauhti.
Autoridad de eſcriptura. tlaneltililoni. tlatolneltililoni.
Autorizada eſcriptura. tlaneltililli. tlatollaneltililli.
Autorizar eſcriptura. ni.tlatolneltilia.na.moxtlatolneltilia
Autoridad tener para hazer algo. ni.nauatile.
Auaricia tener. ni.teoyeuacati. ni.tlatlameti. ni.tzotzoca
teuitzti. ni.tzotzocati. atle niccaualiztlamati. & per meta
phoram. atle niquixcaua. aninocotontlani.

Transcribed PL Corpus

?? ??
Arronarſe la muger, ni no, cuitlapan mauhtia.
Atró nada muger, mocuitlapan mauhtiqui.
Atronar á otro con ruido, ni te nacaztititza,
á tronado allí, tlanacaztitĩztli,
ã[Audiencia de los Juezes. tlacaco ya
coya. tecutlatoloya, 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 aunc̃cuix ça occc97xçac.
Aun no. ayamo.
Aun no ha abuelto, ayatec̃cen casayatec̃ayaíz
tli.
Aun que, yntla nel, ymmanel, ymmaçonel, ii
iuh maciui, maçoiui, maçoneltin, aço yuh.
Aufentat fe. canapa niyauh. nino-yeltia, ni-cl
tlatia. anĩxpa,
Aufencia. canapa yaliztli, netlatiliztli, atẽĩxpa,
Aufente. ayac. canapa yaqui amo ixpa.
Autor hazedor dios. techiuani. tepi quini, te yc
Autoridad de perſona, teixmauhtiliztli. mau
Autoridad tener de perſona, ni tesix mauhti, ni
Autorizada perſona, teixmauhti, teixmamauj
Autoridad de feriptura. tlaneltililoni, tlatolti
Autorizada eſcriptura, tlaneltililli, 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. anĩnocotontlani.



A ante t, & ti. 29

Atronar ſe la muger, nino.cuitlapan mauhtia,
Atronada muger, mocuitlapan mauhtiqui,
Atronar a otro con ruido, nite.nacaztititza,
Atronado aſſi, tlanacaztititztli,
¶ Audiencia delos juezes. tlacacoya.tlatzonte-
coya. tecutlatoloya. teccalli,
Audiencia hazer, nitecutlatoa.nitla,caqui,
Auelo o aguelo, colli tecol.
Aullar. nite,coyoua.
Aullador. tecoyouani.
Aullido. tecoyoualiztli,
Aun. noma. oc noma,
A vna hazerſe, ticcemitoa, tictocentequilia,
A vna parte y a otra, necoc, nenecoc necoccampa occã-
pa ixti.yyuccampaixti.yyuntlapalixti.
Aun aun? cuixçaoc? cuixçac.
Aun no. ayamo.
Aun no has buelto. ayate? cenca ayate? ayalztiuitz? tica
tli,
Aunque, yntlanel, ymmanel, ymmaçonel, immaçonel-
iuh.maciui. maçoiui. maçoneſiui. aço. yuh,
Auſentar ſe. canapa niyauh. nino,yeltia, ni,choloa. nino,
tlatia.anixpa.
Auſencia. canapa yaliztli.netlatiliztli. ateixpa.
Auſente. ayac.canapayaqui amo ixpa.
Autor hazedor dios. techiuani.tepiquini, teyocoyani.
Autoridad de perſona. teixmauhtilizrli. mauizticayotl.
Autoridad tener de perſona. nite,ixmauhti,ni,mauiztic.
Autorizada perſona, teixmauhti.teixmamauhti.
Autoridad de ſcriptura. tlaneltililoni, tlatolneltililoni.
Autorizada eſcriptura. tlaneltililli, tlatollaneltililli.
Autorizar eſcriptura. ni,tlatolneltilia,na,moxtlatolneltilia
Autoridad tener para hazer algo, ni.nauatile.
Auaricia tener, ni,teoyeuacati, ni,tlatlameti.ni,tzotzoca
teuitzti.ni,tzotzocati.atle niccaualiztlamati. & per meta
phoram. atle niquixcaua, anĩnocotontlani.

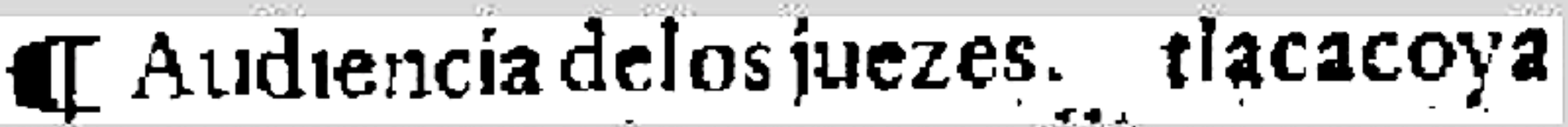
G

Transcribed PL Corpus

Atró nada muger, mocuitlapan mauhtiqui.

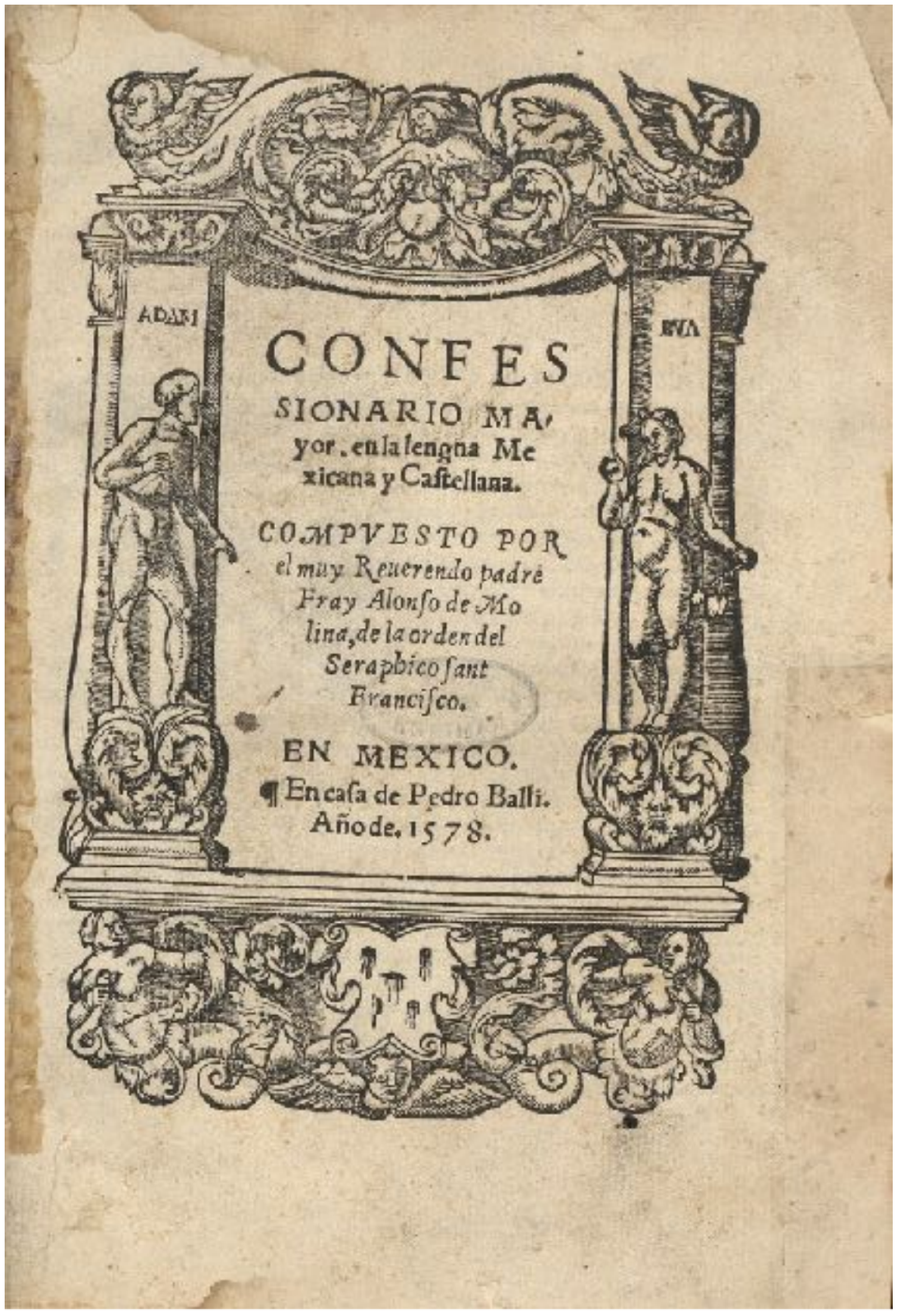
Atronar á otro con ruido, ni te nacaztititza,

á tronado allí, tlanacaztitĩtztli,



ã[Audiencia de los Juezes. tlacaco ya

Transcribed PL Corpus

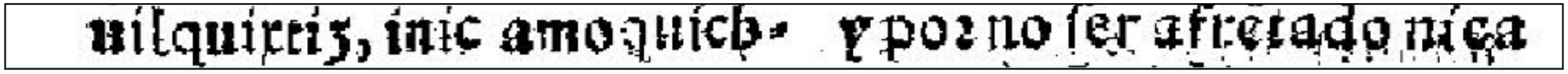


ynic amo mopinaubtiz, ma
nilquiztiz, inic amo quich-
teccatlapiquizque, aubcen
ca mo maubtia, quimacaci
ynuclatzacuiltiloca: çanno
yub motechmonequi, ynic
cenca timomociniz, ipquich
motlapal ticchiuaz inicmu
chi ticcemoz, tiquilnami-
quiz imotlatlacol, inic vel
timo yolcuitiz, inic titlapo
uaz ypantzinco totecuiyo
dios, yuan inipã sacerdo
te, inic amo tipinaubtiloz,
inic amo titlatzacuiltiloz,
çantinanicouaz yuan tic
maquiztizinmanimã. Aub
intlacamo achtopa tiquil-
namiquiz, ticcentlaliz imo
tlatlacol, velnelli yetiqui
tlacoz im moneyolmelaua
liz, amo manelli yeticpale
uiz inmanimã, çan occenca
ye icticcholiniz, amo ma mo-
paleuiloca mochiuaz imo-
neyolcuitiliz, çan mopina
ubtiloca, motelchiualoca
yez, yuã cemicac mixnaua
tiloca mochiuaz. Aub inic
amo tiquitlacoz, mamocen
yollocopa achto xictemo,
yuan xiccentlali ynipquich

nãten q̃ vio ruin cuẽta de si
y por no ser afretado nica
er en algũa falta, y por q̃no
le tẽgã por ladron; y tiene
grã miedo y temor del casti
go que se le podria dar: assi
tienes necessidad de tener
grã solicitud y cuydado, y
vbazer toda su possibilidad
para buscar y pẽsar todos
tus pecados para te confessar ỏ
uidamẽte, y para dar buẽa
cuẽta ỏlãte nro señor dios
y ỏlãte el sacerdote, porq̃
no seas auergõçado y casti
gado, mas ãtes recibas hõ
ra, y alcances la saluacion de
tu aĩa. Y si primero no trux
eres a la memoria tus pecãos
y los recogieres, ciertamẽ
te sera tu cõfessiõ inualida,
falta y no verdadera, ni me
nos sera vtil y prouechosa a
tu aĩa, ãtes la afligiras mu
chomas, y no sera para tu
fauor la tal cõfession mas
para tu cõfusion, y cõdena
cion y para q̃ eternalmẽte se
as ỏsechado. Y porq̃ no sea
imperfecta la tal cõfessiõ tu
ya, busca primero ỏ todo tu
corazõ y ayũtarẽ uno todos

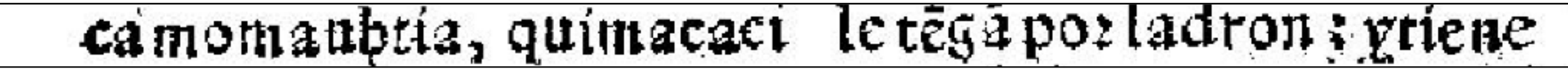
Transcribed PL Corpus

ynic amo mopĩnãuhtij, ma nãten. q̃Dior uincuẽtaoe-lĩ



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

teccatlapiquijque, auh cen eren algua falta, y poz q̃n.o



ca momaubtia, quimacaci le tẽgã poz ladron ; ytiene

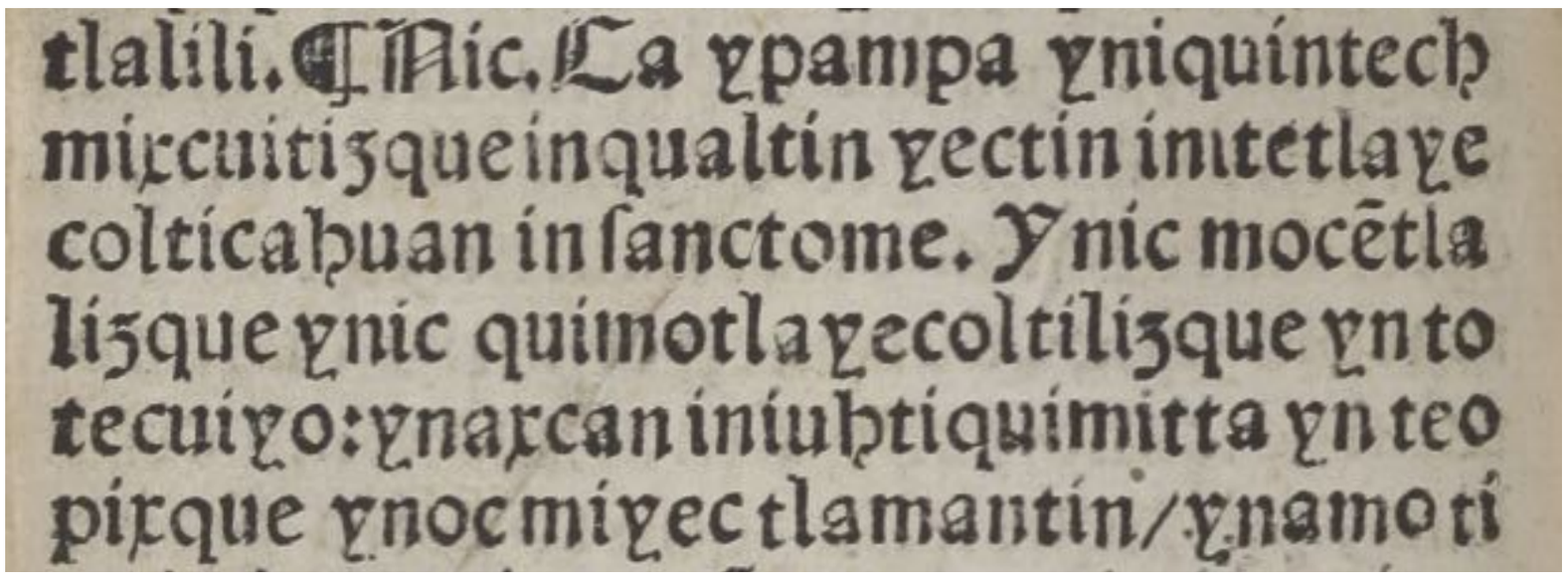
Evaluating OCR

Google

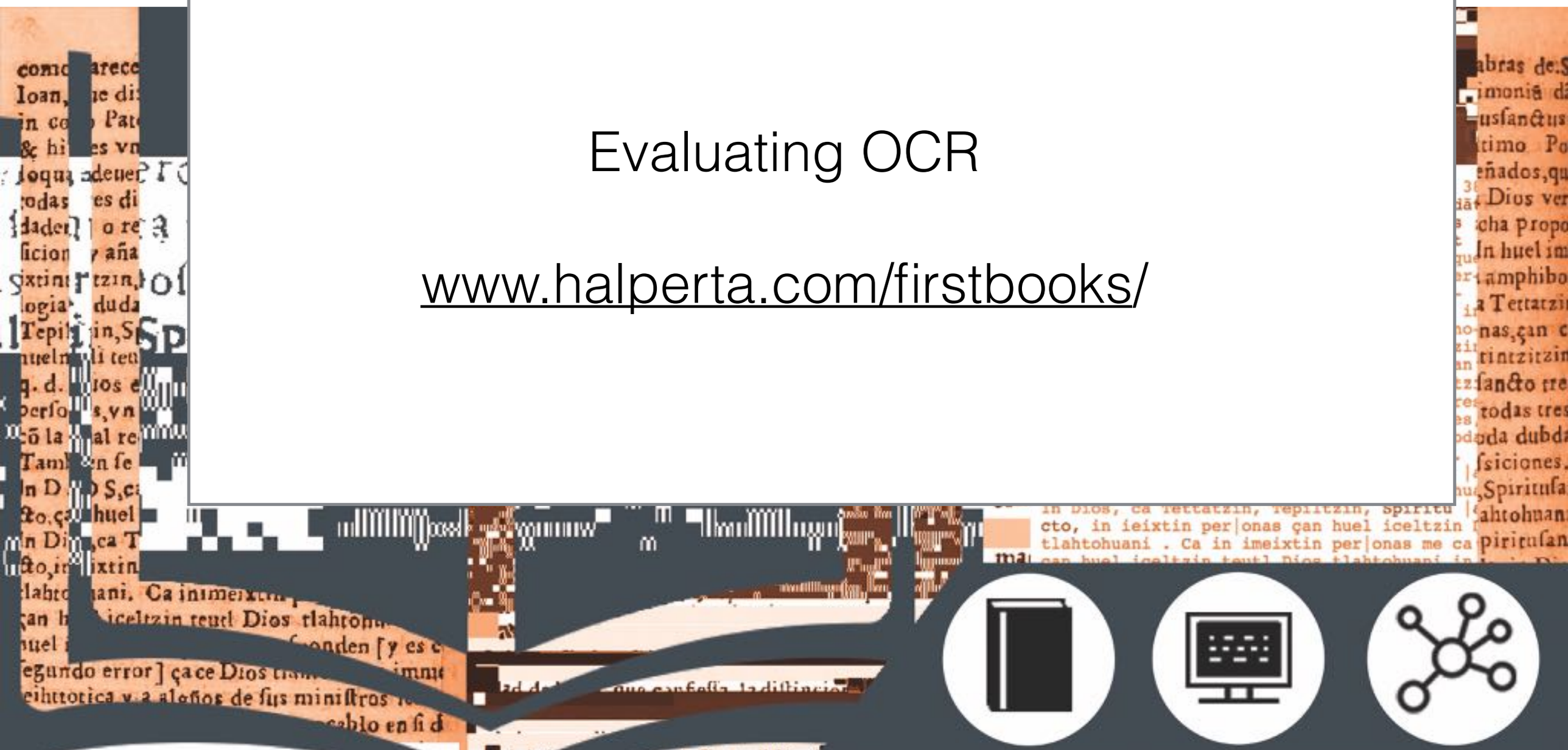
tlalili. Clic, La epanipa yniquíntech mircuitizqueinqualtin pectin initetlaye colticahuan in fanctome. Ynic mocētla lisque ynic quimotlagecoltilisque ento tecuiyo:ynarcan iníuhtiquimitta yn teo pirque ynocmigec tlamantin/znamoti

Ocular

lalili. di. il lic, ĩCa ypampa yniquĩntech mixcuitizque in qualtin yẽctin in ĩtẽtlaye colticahuan in iancto me. Ynic mocẽtla lizque ynic quimotlayecoltilĩzque yn to tẽcuiyo: yn axcan in iuh tiquimitta yn teo pixque yn oc miyec tlamantin ? yn amo ti


tlalili. Clic. La ypampa yniquintech
mixcuitizqueinqualtin yectin initetlaye
colticahuan in ſanctome. Ynic mocētla
lizque ynic quimotlayecoltilizque yn to
tecuiyo:ynaxcan iniuhtiquimitta yn teo
pixque ynocmiyec tlamantin/ynamo ti

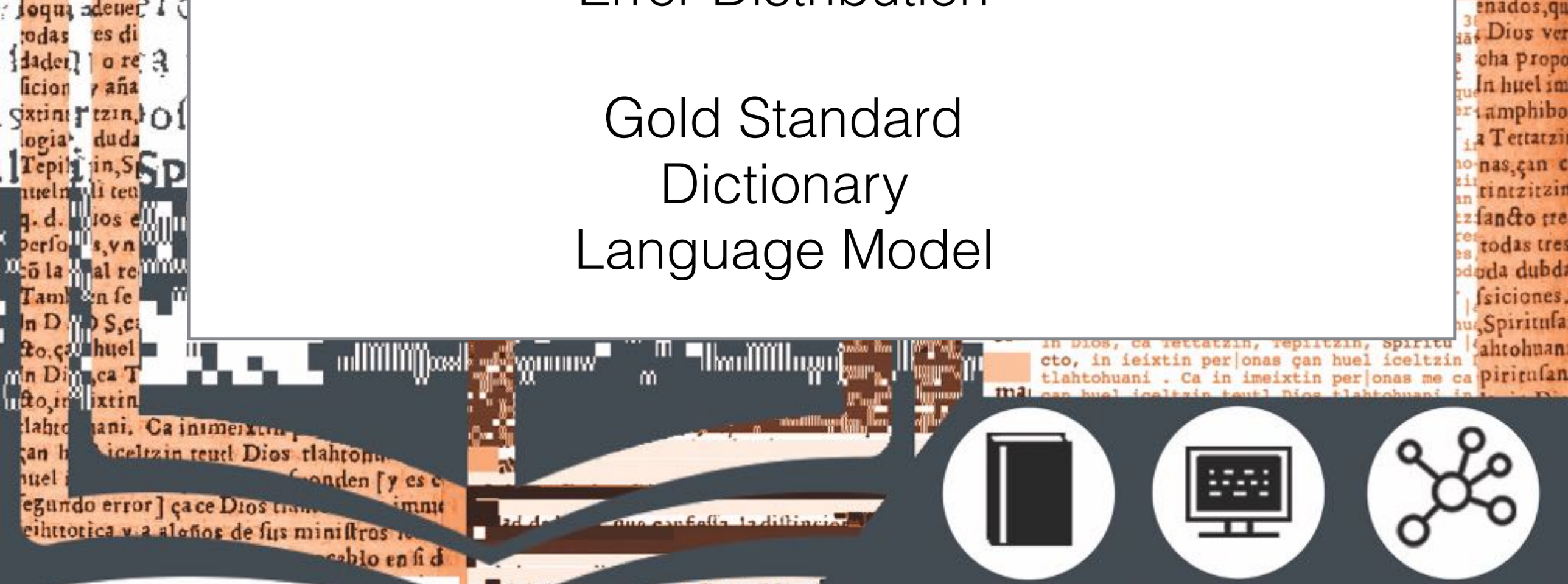
Evaluating OCR



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

Exercise:
Transcribing Colonial Documents with From the Page

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

Collections

Joaquín Garcia Icazbalceta Collection

Lista de los pueblos de indios…

https://fromthepage.lib.utexas.edu/display/read_work?document_set_id=8&work_id=564

Instructions

Cities: [[cityname]]

Example: [[Oaxaca]]

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