

Transcribing Multilingual Documents in the Digital Age

Hannah Alpert-Abrams
University of Texas at Austin



Part 1: Transcription Options

Part 2: Transcription Problems

¶ **A**lic. ca yampam iinic yehuantin quicui
lozque ininemiliztin ynictlalticpacmo
neantico in totecuyó Jesu xpo: intleyñ
quimochihuilico in quimotemachtilico:
inictehuantin tictotepotztoquili que y
nitemachtiltsin. ¶ **N**o. Catlehuatl ino
quicuiloque in nahtuintin in Euangeli-
tame.

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

¶ **N**o. nicmatiznequi ynictiazque ynil-
huicac cuiꝝ ga yeyro ticneltocazque yn
oquicuilotiaque.

¶ **A**lic. ca itlacamo tictopielicā tictone
milizticā in teotenuahuatiltzin, y yehua-
tin in quicauhtiaque caniman ahuel tiaz
que ynilhuicac.

¶ 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.

Accessibility
Discoverability
Preservation
Research

Labor
Cost
Skill
Accuracy
Privacy

Crowdsourcing

Automatic Transcription

**JOIN US!****LEARN HOW TO TRANSCRIBE**

Become a Smithsonian Digital Volunteer and help us make historical documents and biodiversity data more accessible.

Join 8,650 volunteers to add more to the total 287,436 pages of field notes, diaries, ledgers, logbooks, currency proof sheets, photo albums, manuscripts, biodiversity specimens labels that have

BROWSE PROJECTS

Select a category below to begin browsing projects.

Select a Category ▾

DISCOVER THE FREEDMEN'S BUREAU

**North Carolina Assistant
Commissioner, Registers of Letters**
View Project

LATEST UPDATES

mandc transcribed a page from Charles Francis Hall Journal with Navigational Notes 1861

sagehen transcribed a page from North Carolina Assistant Commissioner, Endorsements Sent, Vol. 3 (20), Mar. 4-Dec. 30, 1868

Katies marked for review a page from Charles Lang Freer's letters to Frank

Author: Packe,
Susanna, fl. 1674.

Title: Cookbook of
Susanna Packe
(manuscript).

Catalogue

Original

A. Transcribe Text

Identify Graphic

Tutorial

Remember, you're not
required to transcribe
the whole page!

I'm done!

A

for a soote of the lippes.

2 spoonfull of yle of sweetabacione.

then 3 spoonfulls of cold ffectuate.

fat of recce rage water 2.ounces of

sugardandy beaten fine mynded al

shear rogered and beat it one hower with a

boon fech it be very wight then take this

for a drak of Hickory shooles

ake that is new borne or any one whilke

spaf coagd or borselotis also very

good for the bone.

gasp of sparslan is red to be a
paine remedy for a cough and
consumpcion.

wounds unquinton sumptation

i yere to dispeche any swelling ffor it bee

boyled and booke & bound together

so i yle of camomile.

OVNTMENS

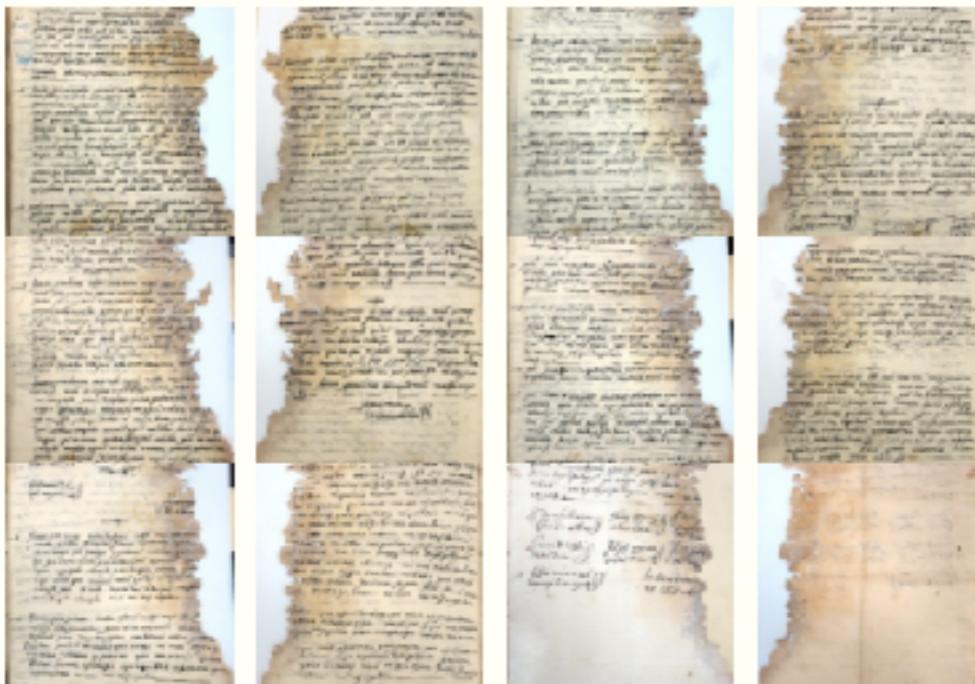
1674

Sample Text

Axcan ipan miercoles çempuali macultonal metztli enero x-

Created

21-04-2016

Files**Transcribe This Item**

1. [AGCA A1 Leg6071 Exp 54891-1.JPG](#)
2. [AGCA A1 Leg6071 Exp 54891-2.JPG](#)

Crowdsourcing

Accurate
Affordable
Needs Volunteers
Needs Infrastructure
Needs Maintenance

scripto

a community transcription tool

[Home](#) [Showcase](#) [User's Guide](#) [Download](#) [News](#) [About](#)

A free, open source tool enabling community transcriptions of document and multimedia files

Scripto brings the power of MediaWiki to your collections. Designed to allow members of the public to transcribe a range of different kinds of files, Scripto will increase your content's findability while building your user community through active engagement.

[Download](#)

[Learn More](#)

difficulties, incurred a very different fate; mine is the unhappy station, in which I must hear complaints, without having it in my power to redress the grievances?"

"September 30^d 1782.
"The moment you have taken your determination, what

The moment you have

there will be no doubt, in case Charlton should be wanted indeed, we had almost better give any price, than think of sending it from here. We have met with so many losses by delays, that we have little hope of success, should it be again attempted; however, if the clothing cannot be had with you, must go from hence."

"November 5^d 1782.

<http://scripto.org/>



“Simply the finest crowdsourcing manuscript transcription software on the planet”

That's a quote by Max Spiegel, who runs Zeprapedia, Philip K. Dick's Exegesis, at Penn State. He's one of many people running collaborative transcription projects on FromThePage.

You've scanned your collection for years, but no one can find the text locked in your images. Use FromThePage to free your images and engage collaborators everywhere.

From the Page

Crowdsourcing

Trauma

Labor

Crowdsourcing

Automatic Transcription

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

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

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

¶ Nic. ca itlacamo tictopielicá tictone milizticán initeotenahuatiltzin, y yehuá tin inquicauhquiaque caniman ahuel tiaz que yn ilhuicac.

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

¶ 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 cuiotiaque çan nim~a ahueltitomaquiax tizqui.

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

¶ Nic. ca ~itlacamo tictopielic~a tictone milizticán initeotenahuatiltzin, y yehu~a tin inquicauhquiaque caniman ahuel tiaz que yn ilhuicac.

Chic. 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.

Chic. 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.

CMo. nicmatiznequi ynictiazque ynil huicac cuiꝝ ga yeyyo ticneltocazque yn oquicuilotiaque.

Chic. ca itlacamo tictopielicā tictone milizticā in teotenahtuatižtin, y yehuá tin in quicauhtiaque caniman ahueltiaz que ynil 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-

Options for Automatic Transcription

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](#)[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

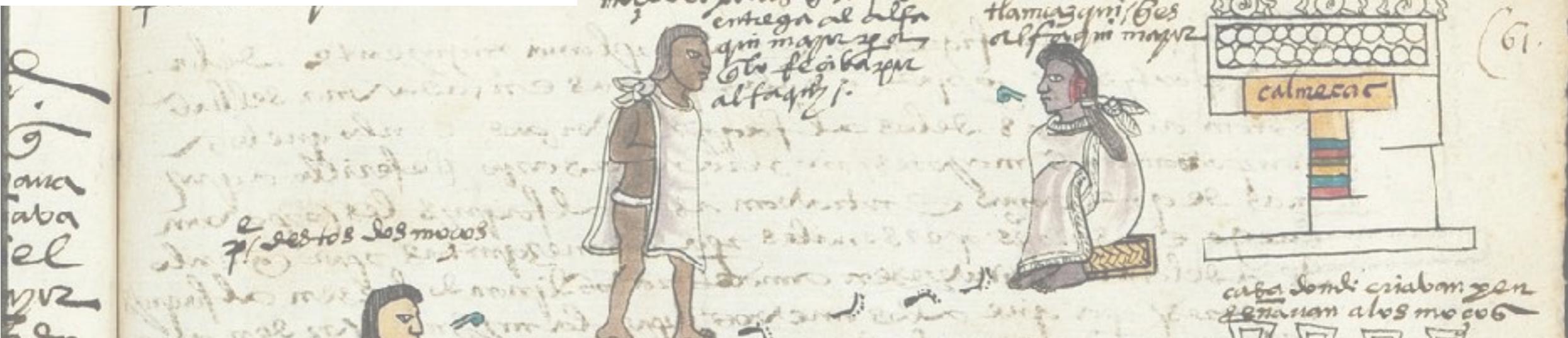
Technologically complex Command-line operations Single-document transcription

Ocular

Part 1: Transcription Options

Part 2: Transcription Problems

Colonial Transcription



The Codex Mendoza

Tlacuilo: Francisco Gualpuyogualcal

Transcription: Juan González



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 desencion: exor-
tandolos ahuyr, losvi-
cios, yaque se diesen,
ales exercicios denoble
za, y de virtud.

Iccax tolli omome ca-
pitulo, vncan motereoaa
centhamati circa qualli, le
nonotsaliztlatolli, necemij
listiloz, injc qujn nonotsa-
ja, i pilhoan tlatoanj: inj
quacie ixtlamati, ietlaca
que, qujn tlaguauh mo-
raio, injc qujtlacahujs
que, injc qujch in aqualli,
mnaectli. Auh injc qujtl
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

bieron³ S. principales, que
gobernaban⁴ por capitanes. Los de
Culiba aparecieron⁵ 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 Culancion⁶ Pueblo
^{Culaniungs} Tlanco⁷ Tlancio
tollan (hoy Tula)⁸ Tlancio
fueron a Tula, doce leguas de Mexico, á la
parte del Norte, y vinieron poblando hacia
Texcoco Texcoco
Tlancio, que es la orilla del agua de la La-
guna de Mexico, cinco leguas de Tlancio
y ocho de Tlancio. Tlancio está la parte de
Oriente, y Mexico al Occidente, la laguna
en medio. Algunos quieren decir que Te:
xoco se dice Culiba por respeto de estos
que allí poblaron. Despues el Señorío de
Texoco Texcoco

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-²⁷
num Indorum, Indis, quod dicitur Octli :&
etiam de iis qui eis venduntur non verum His

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

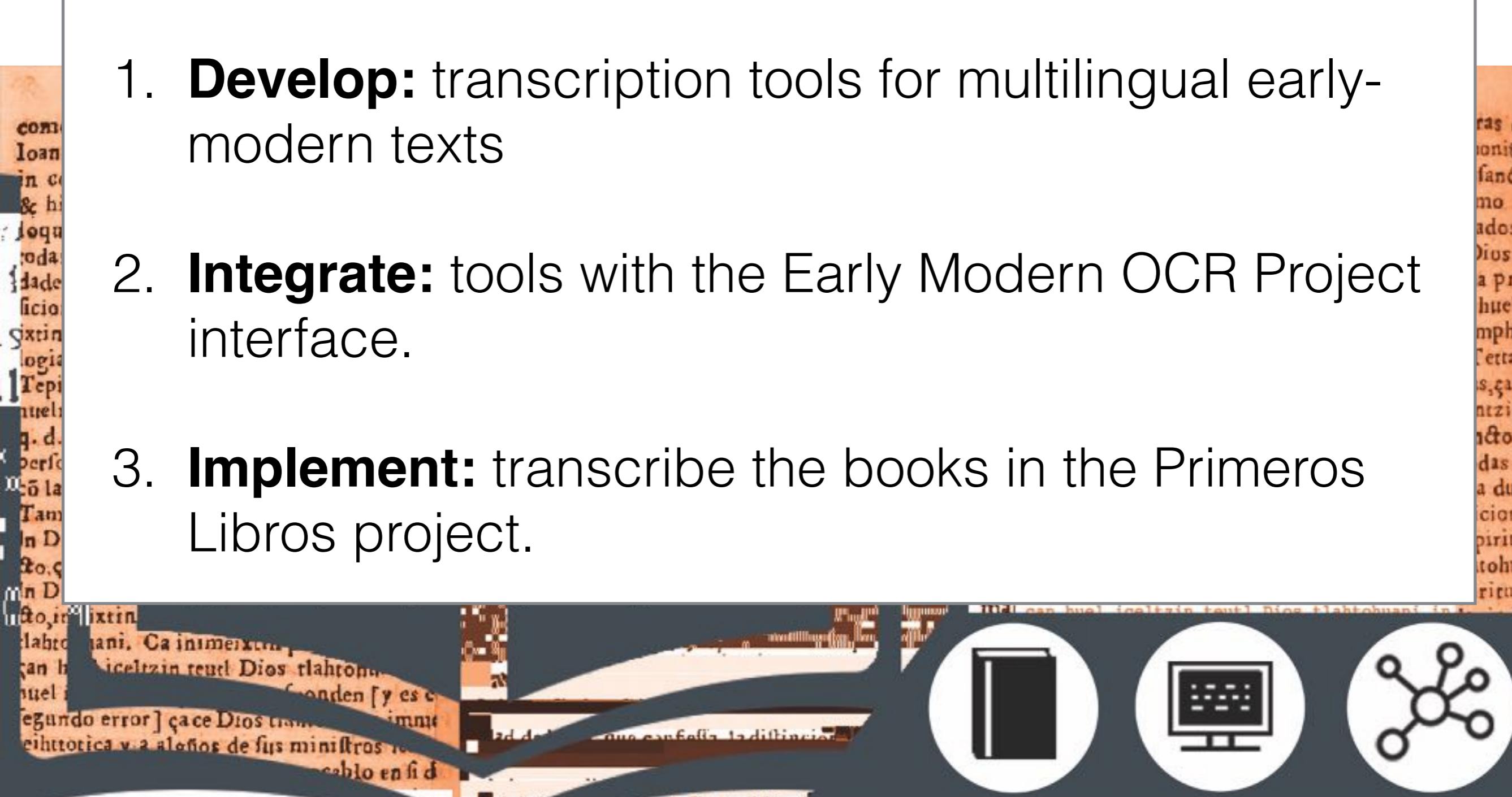
CAPIT. V. VN CAN M
toa, motenua, in micētlamantli tl
matiliztli in itechca Anima, iuhqui
ma micētlamantli xochiqualqua
yanitl, icenca vel ita

“^{et quo} 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.



How does Ocular work?

replitzin, Spiritu |ancto, ei per|onas, çan tintzitzim
huel nelli teutl Dios in huel imeixtintzitz |ancto tre
q. d. Dios es Padre, Hijo, y Sp̄o |ancto tre
per|onas, vn |clo Dios verdadero todas tres |odas tres
cō la qual reduplicacion |e quita toda dubdoda dubda
También |e quita con e|tás propo|siciones.
In DTO S.ca Tettatzin, Tepiltzin, Spiritu |
tto . çan huel iceltzin teutl Dios tlahtohu, Spiritus
In Dios, cá Tettatzin, Tepiltzin, Spiritu |
cto, in ieixtin per|onas çan huel iceltzin
tlahtohuani . Ca in imeixtin per|onas me ca piritus an
çan huel iceltzin teutl Dios tlahtohuani in ieixtin
tlahtohuani. Ca inimeixtin
gan huel iceltzin teutl Dios tlahtohuani
huel iceltzin teutl Dios tlahtohuani
segundo error] çace Dios tristeza immut
cihttatica v. a años de sus ministros re
cable en si d



Challenges of Early Modern OCR

Wandering baseline

§. Los verbales en liztli, significan

Uneven inking

Positionem completuam. Y en la d. 16 q.

Unfamiliar Typefaces

Moctezuma tley pampa ynoquim



Challenges of Early Modern OCR

praesertim urgente causa

ligature non-standard spelling diacritic character ellision obsolete character

praesertim urgente causa

segundo error] çace Dios tráe immut
cihtotica v. a años de sus ministros re
cable en si d



Solutions to Early Modern OCR

Font Model

+

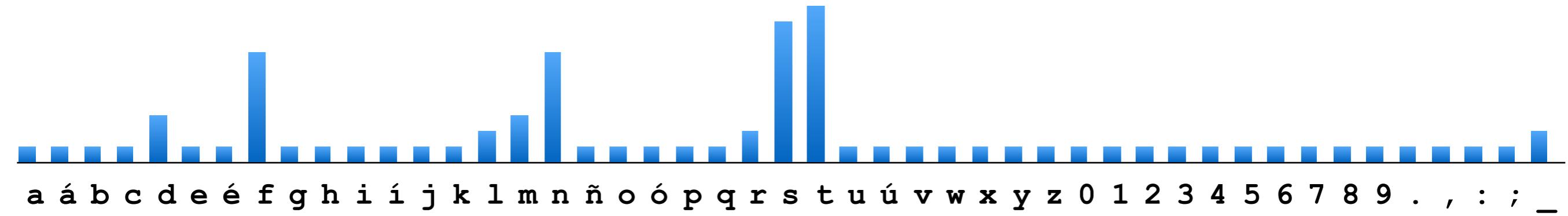
Language Model

Font Model

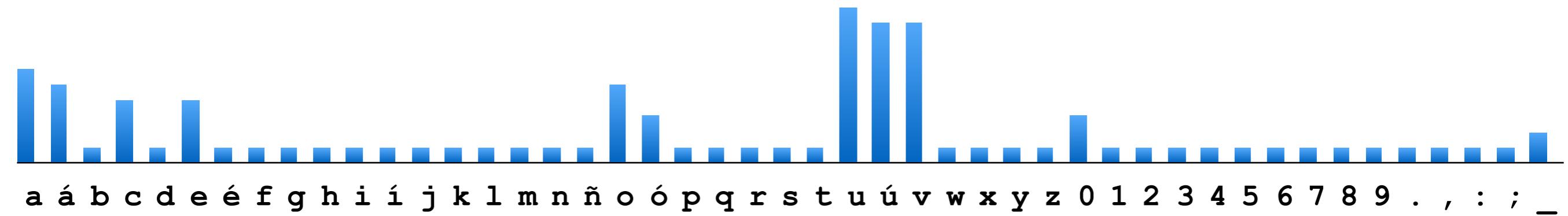
í

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 . , : ;

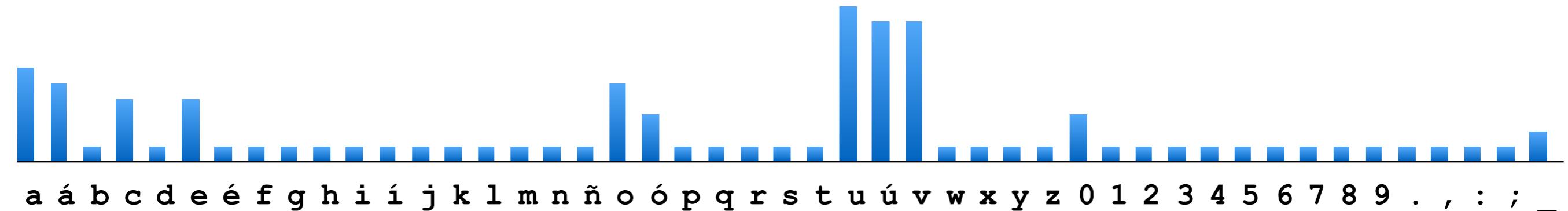
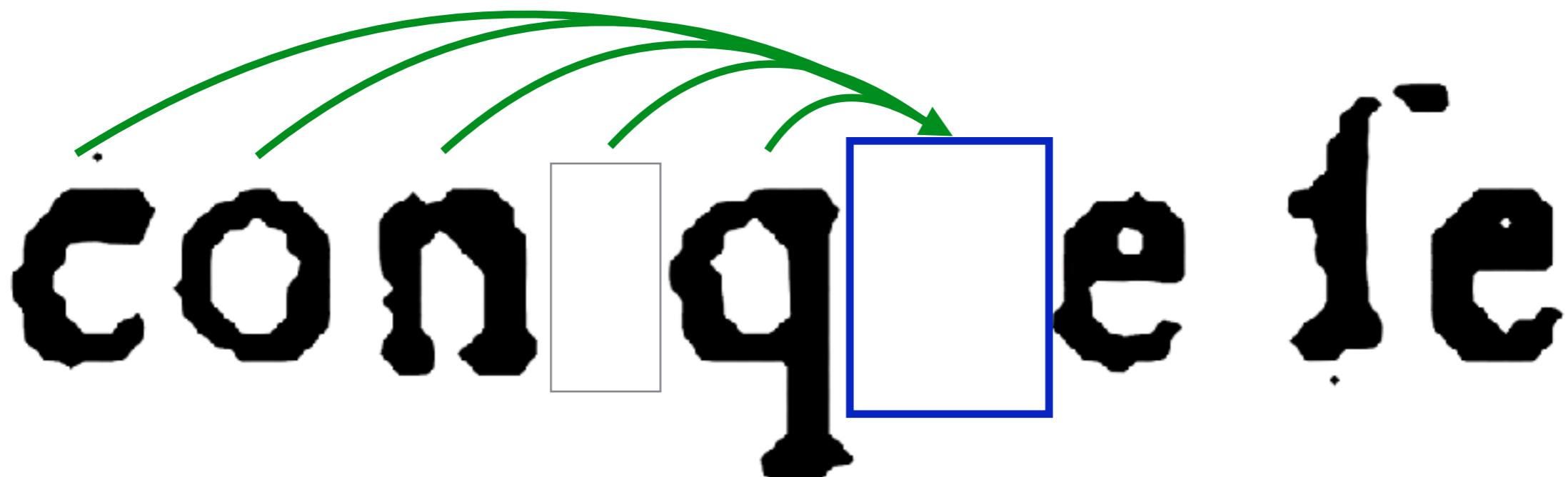
Language Model



Font Model



Language Model



Language Model

con que te

con que fe

"¶ 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-²⁷
num Indorum, Indis, quod dicitur Octli :&
etiam de iis qui eis vendunt vinum verum His

- Normalized text: illustres. E logo os Reitores **que** taõ juntos, derão o mesmo
Historical text: illustres. E logo os Reitores \~qe taõ juntos, derão o mesmo
- Normalized text: aviso a seus Collegios, **sendo** em todos, & cada hum delles
Historical text: avifo a feus Collegios, fendo em todos, & cada hum delles

illustres. E logo os Reitores \~q eraõ juntos, derão o mesmo
aviso a seus Collegios, sendor em todos, & cada hum delles

Ocular Interface for eMOP

Chrome File Edit View History Bookmarks People Window Help

19% Mon 4:20 PM halperta C

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... n	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 dh-db02.tamu.edu/corpus-manager/document-details/?corpus-id=7&document-id=1056

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
ID: 134355 Image: pl_tamu_012_00003-1000.jpg
fb_page_id: pl_tamu_012_00003

pl_tamu_012-rincon_OCR_5-26-

Ocular Comparison: pl_tamu_012_00003-1000_comparisons.txt

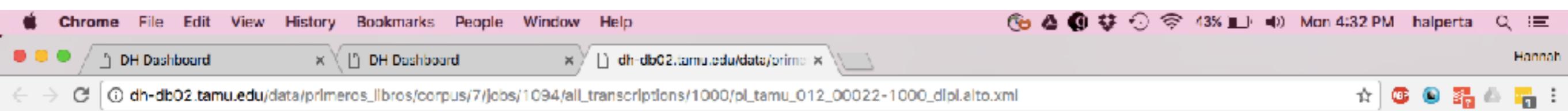
Ocular Interface for eMOP

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... 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-
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>
```

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, aninococontlani.

G

Transcribed PL Corpus

?? ??

Arronarfe la muger, ni no, cuitlapan mauhtia.
 Atró nada muger, mocuitlapan mauhtiqui.
 Atronar á otro con ruido, ni te nacaztitza,
 á tronado allí, tlanacaztitítztlí,
 [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-
 Aulido. tecoyoualiztli,
 Aun. Jsoma, oc noma.
 A yn a hazerfe, 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. aníxpa,
 Aufencia. canapa yaliztli, netlatiliztli, atéixpa,
 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. tlaneltililoni, tlatolti
 Autorizada ecriptura, 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. anñnocotontlani.

A ante f, & d. 29

Atronar se la muger, nino.cuitlapan mauhtia,
 Atronada muger, mocuitlapan mauhtiqui,
 Atronara otro con ruido, nite.nacaztitza,
 Atronado assi, tlanacaztitítztlí,
 [Audiencia delos juezes. tlacacoya.tlatzonte-
 coya. tecutlatoloya. teccalli,
 Audiencia hazer, nitecutlatoa.nitla.caqui,
 Auelo o aguelo, colli tecol.
 Aullar, nite,coycua.
 Aullador, tecoyouani.
 Aulido, tecoyoualiztli,
 Aun, noma, oc noma,
 A vna hazerfe, 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 buelto, 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,aníxpa.
 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, tlaneltililoni, tlatolneltililoni.
 Autorizada ecriptura, tlaneltililli, tlatollaneltililli.
 Autorizar ecriptura, ni,tlatolneltilia,na,moxtlatolneltilia
 Autoridad tener para hazer algo, ni,nauatile.
 Auaticia tener, ni,teoyeuacati, ni, tlatlameti, ni, tzotzoca
 teuitzti,ni, tzotzocati, atle niccaualiztlamati, & permets
 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,

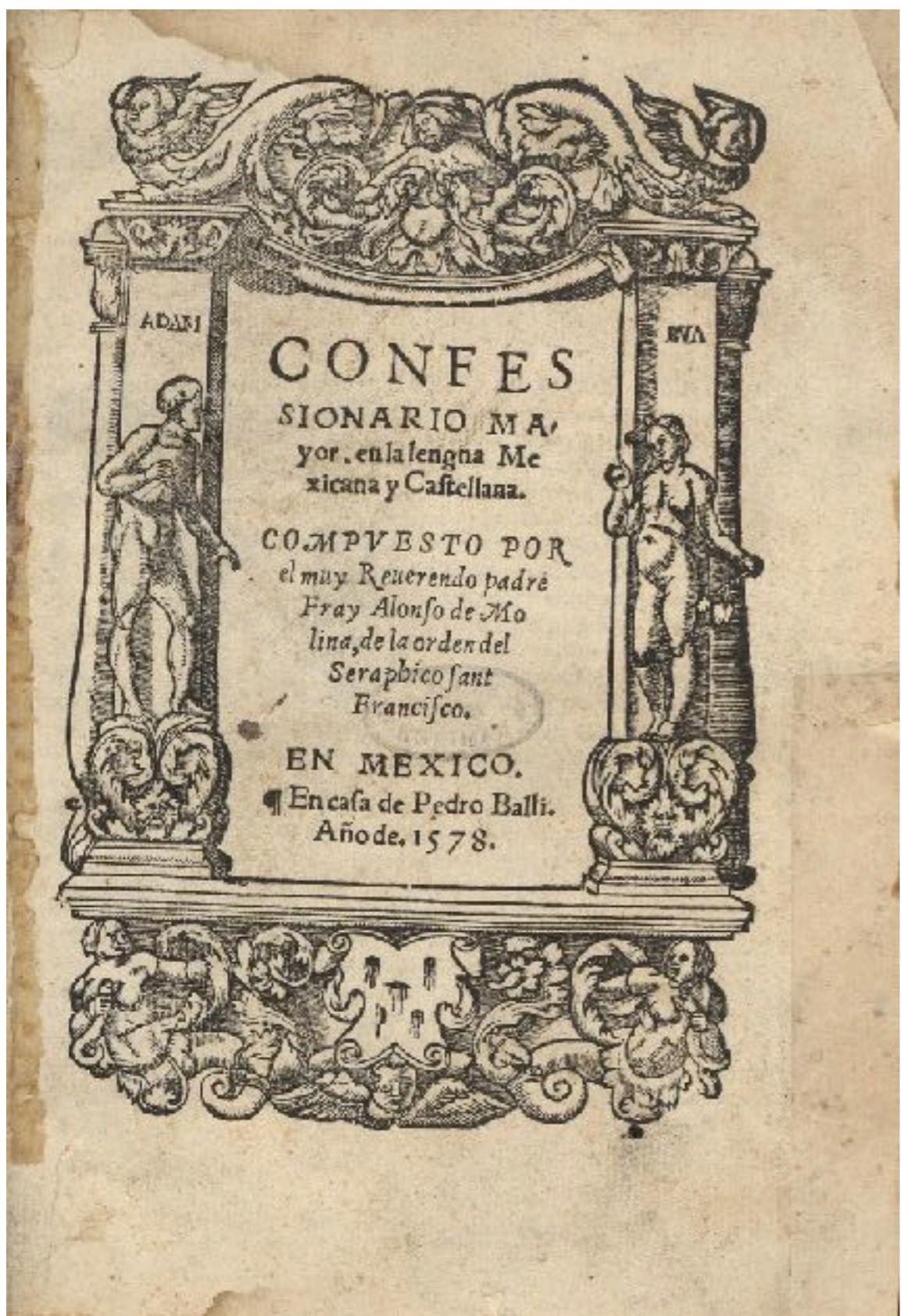
Atronado así, tlanacaztititztli.

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 ticchiuaç inicmu
chi tictemoz, tiquilnami-
quizimotlatlacol, inic vel
timogolcuitiz, inic titlapo
naz yxpantzinco totechiyo
dios, guaninixpá sacerdo
te, inic amo tipinauhtloz,
inic amo titlatzacuiltyloz,
çantinaniçonaç ywan ric
maquixtizimanimá. Ahu
intlacamo achtopa, tiquil-
namiñiz, tictentlaliz imo
tlatlacol, vel nelli yctiqui
tlacozi monegolmelaua
liç, amoma nelli yeticpale
uiçimanimá, çan occenca
ye ictictoliniz, amoma mo-
paleuiloa mochiuaç imo
neyolcuitiliz, çan mopina
ubtloca, morelchiualoca
yez, yuacemicacmitnaua
tiloca mochiuaç. Ahu inic
amotiquitlacoz, mamocen
yollocopa achto xictemo,
guan riccentlali ypixquich
uaten qdioruincueta desí
y por no ser afretado nica
er en algúia falta, y por qno
le tégá por ladron; y tiene
grámiedoy temor del casti
go que se le podria dar; así
tienes necesidad de tener
grásolicitud y cuidado, y
obazertodatu posibilidad
para buscar y pésar todos
tus pedos para te confessar d
uidamente, y para dar buéa
cuéta dlate nroseñor dios
y dlate el sacerdote, porq
no seas auergoçado y casti
gado, mas átes recibas ho
ra, y alcáces la saluació de
tu aia. Y si pmero no truje
res ala memoria tus pedos
y los recogieres, ciertame
te sera tu cōfesió invalida,
falta y no veradadera, ni me
nos sera útil y puecposa a
tu aia, átes la astigiras mis
chomas, y nosera para tu
fauor la tal cōfession mas
para tu cōfesion, y cōdena
ció y para q eternamente se
as desechado. Y porq nosea
isperfecta la tal cōfesió 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

Transcribed PL Corpus

See more sample transcriptions:

halperta.com/firstbooks

Part 1: Transcription Options

Part 2: Transcription Problems

Part 2.5: Ocular Problems

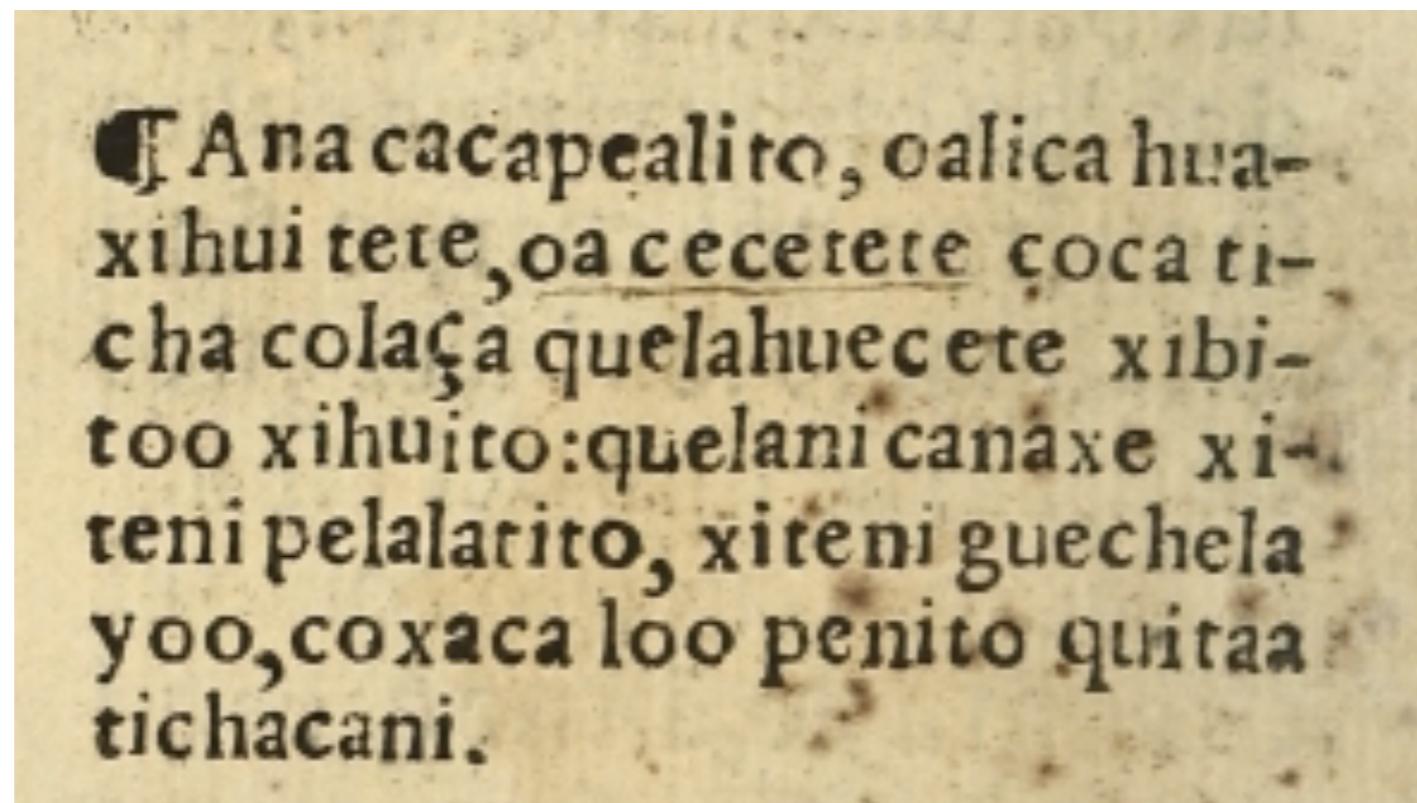
spacing

æ Ana **caca pealito**, oalica hua-
xihui teresoa cgçetete goca ti-
cha colaça quelahue cete Xibit-
too xihui torquelani canâ xe xi-
teni pelala rito, xiteni Sue chela
yoo, coxaca loo pẽniso guitia

dirty OCR

96956894-

overcorrection



morpheme
parsing

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: Iglesia. in toquichtli xipapaqui, ximotla

teoyotica
+
tawewetl

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

over -
correction

Norm: *l̄ yollo, oquimopanatili* in cemana-
Hist: *l̄ yollo, oquimmopanauili* in cemana-

false
substitution

Mod: oac tlaca.
Hist: oac tlaca.

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

correct
substitution!

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

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

Reading the First Books Future:

Complete transcriptions
Searchable interface
Downloadable corpus
Diplomatic digital editions
Corpus analytics

sites.utexas.edu/firstbooks/



30 MAYO 1978
La Tardé

Página 4

Guatemala, 30 de mayo de 1978

LA TARDÉ

PROBLEMAS SOCIALES

DE LA NOCHE A LA MAÑANA:

20 FAMILIAS FUERON DESALOJADAS DE SU VIVIENDA

■ LOS TRACTORES Y HOMBRES ARMADOS DE MACHETES LAS SACARON DE LOS PREDIOS CEDIDOS EN 1966



De la noche a la mañana, sin ninguna explicación, veinte familias —de las 575 que se han instalado— que construyeron su vivienda en un predio que les cedió la municipalidad en 1966 en las inmediaciones de la avenida del Ferrocarril y 27 calle de la zona 8, se vieron frente a tractores y hombres armados de machetes de una empresa privada, que procedieron a desalojarlas.

Inexplicablemente, a menos que la municipalidad capitulina



Página 4 Guatemala, 30 de mayo de 1978 IATARDE PROBLEMAS SOCIALES DE LA NOCHE A LA MAÑANA 20 FAMILIAS FUERON DESALOJADAS DE SU VIVIENDA O Los TRACTORES Y HOMBRES ARMADOS DE MACHETES LAS SACARON DE LOS PREDIOS CEDIDOS EN 1966.

han permitido el paso para defender sus derechos amenazándolos con machetes y diciéndoles que "si entran con todo y los niños los van a moler con los tractores".

"Nosotros tenemos órdenes de jarzarios de aquí", les

a invasión de los tractores de una compañía privada que se apoderó del terreno que vienen ocupando desde hace 12 años.

DESTINADO PARA UNA GUARDERIA.

pasado.

Posteriormente estuvo una señora perdió y "los vecinos tuvieron que sepultarla" ya que no había ninguna ayuda de la municipalidad.

"ARREGLENLO

N C I A