

Reading the First Books

Automatic transcription for multilingual, early modern printed documents

Hannah Alpert-Abrams

Reading the First Books

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

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

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

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

University of Texas Libraries: Aaron Choate

Chic. ca xpampa inic yehuantin quicui
lozque ininemiztzin ynictlalticpac mo
nemitico in totecuyo Jesu xpo: intley n
quimochihuilico in quimotemachtilico:
inicttehuantin tictotepotztoquilizque y
nitemachtiltzin. **C**Mo. Catlehuatl ino
quicui loque in nahuintin in Euangelis-
tame.

Chic. Ca yehuatl ynauhtlamantli
yn Euangelio ynipan mopia ynixquich
totech monequi in ticneltocazque yn tic
chihuazq. Yhua intley n tictlalcabui3q
intictelchihuazque. Yhuan intley tichi
huazq yye yrqch totech monequi. Ah
yntlacamo ticneltocazque in tlein qui-
cuilotiaque can nimā ahueltitomaquix-
tizqui.

CMo. nicmatiznequi ynictiazque yn il-
huicac cuix ca yeyyo ticneltocazque yn
oquicuilotiaque.

Chic. ca itlacamo tictopielicā tictone
miliztican initeotenahuatiltzin, y yehua
tin inquicaubtiaque caniman ahuel tiaz
que yn ilhuicac.

christiana. Fo 10.

¶ Nic. ca ypampa inic yehuantin quicui
 lozque ininemiztzin ynic tlalticpac mo
 nemitico in totecuyo Jesu xpo: intle yn
 quimochihuilico in quimotemachtilico:
 in ictehuantin tictotepotztoquilizque y
 nitemachtiltzin. ¶ No. Catlehuatl ino
 quicuiloque in nahuantin in Euangelis-
 tame.

¶ Nic. Ca yehuatl yn auhtlamantli
 yn Euangelio yn ipan mopia yn ixquich
 totech monequi in ticneltocazque yn tic
 chihuazq̃. Yhuā intle yn tictlalcahuizq̃
 in tictelchihuazque. Yhuan intle y tichi
 huazq̃ yye yxq̃ch totech monequi. Tluh
 yntlacamo ticneltocazque in tlein qui-
 cuilotiaque çan nimā ahueltitomaquix-
 tizqui.

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

¶ Nic. ca itlacamo tictopielicā tictone
 miliztican in iteotenahuatiltzin, y yehuā
 tin inquicauhtiaque caniman ahuel tiaz
 que yn ilhuicac.

b ij

christiana. Fo 10.

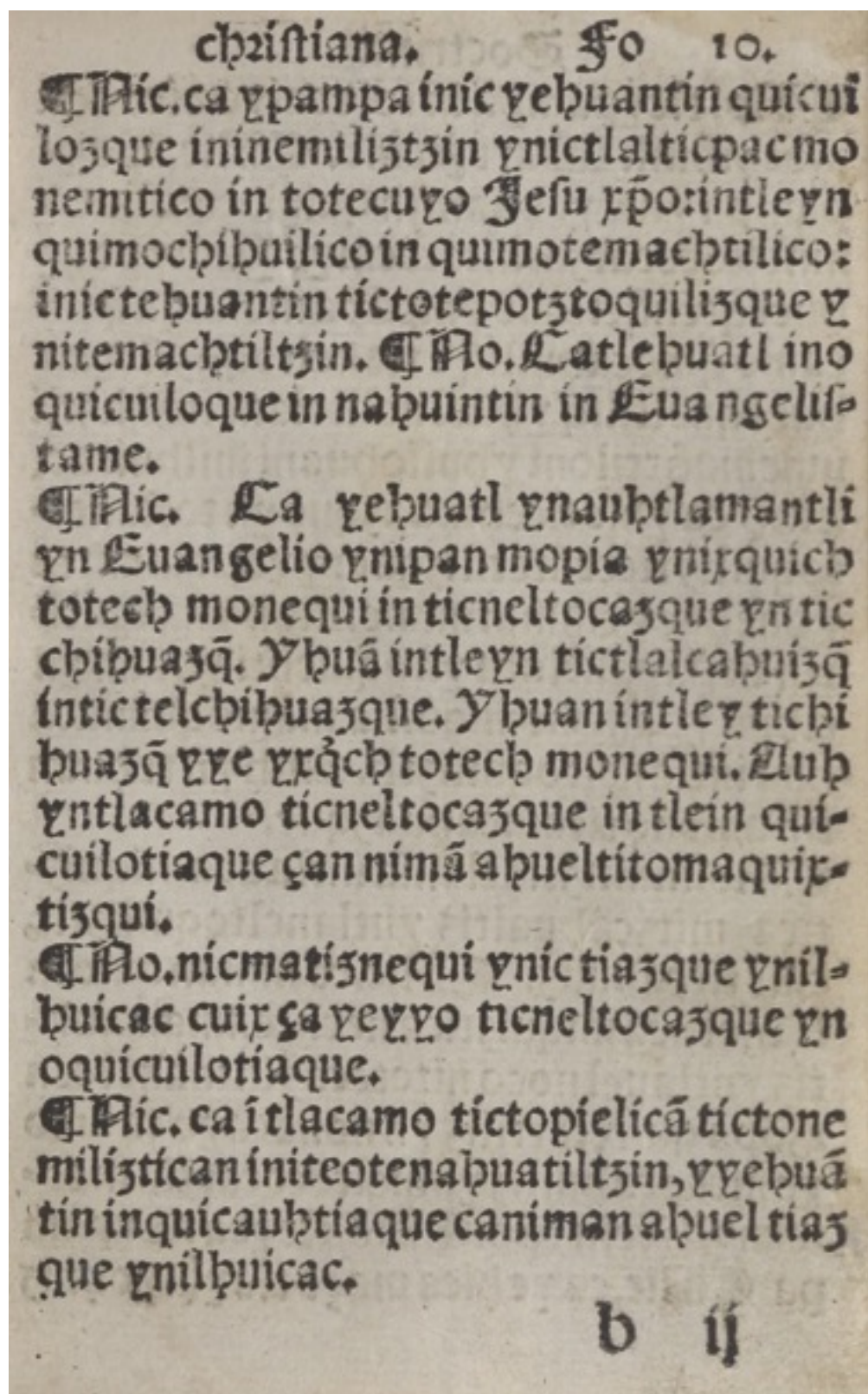
¶ Nic. ca ypampa inic yehuantin quicui
 lozque ininemiztzin yn ic tlalticpac mo
 nemitico in totecuyo Jesu x~po: intle yn
 quimochihuilico in quimotemachtilico:
 in ictehuantin tictotepotztoquilizque y
 nitemachtiltzin. ¶ No. Catlehuatl ino
 quicuiloque in nahuantin in Euangeli-
 tame.

¶ Nic. Ca yehuatl yn auhtlamantli
 yn Euangelio yn ipan mopia yn ixquich
 totech monequi in ticneltocazque yn tic
 chihuaz~q. Yhu~a intle yn tictlalcahuiz~q
 in tictelchihuazque. Yhuan intle y 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 in iteotenahuatiltzin, y yehu~a
 tin inquicauhtiaque caniman ahuel tiaz
 que yn ilhuicac.

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 sovoovosovovovo>ovo r**
r^r^i^i^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- «^lo r-^oo Oso — r^
vo SO vO \0 vO *0 vO vO nO so \0
sOvO vososOnOvOvOvosOvOvO

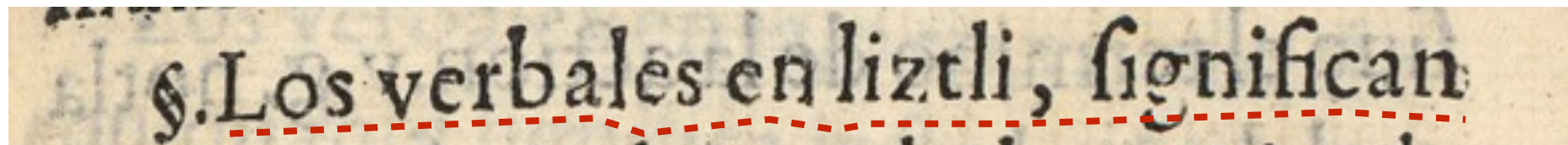
00 (7s O — «* < ' ♦ N ^ w^sO r-»00
(7s o «- n r^ 't-

Ocular: Historical Document Recognition System

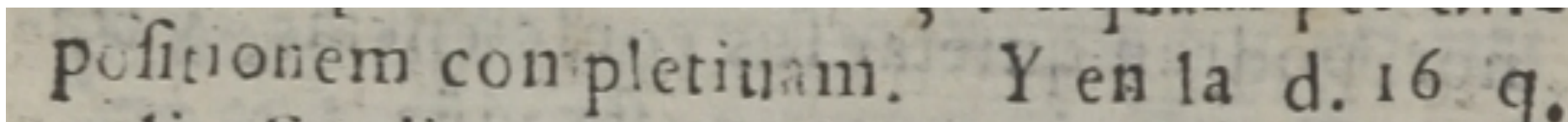
Taylor Berg-Kirkpatrick et al., UC Berkeley (2013)

1. Material Analysis: Font Model
2. Linguistic Analysis: Language Model

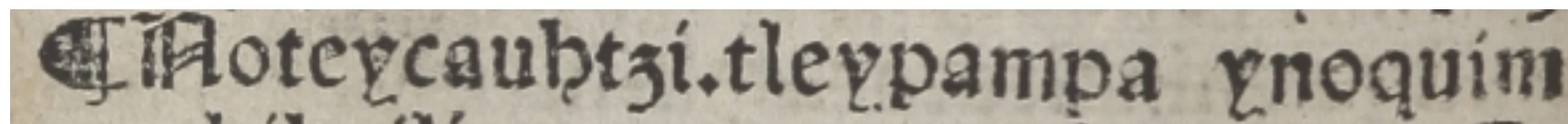
Wandering baseline



Uneven inking



Unfamiliar Typefaces



Variable Orthography

præferunt vrgēte causa

Multilingual Text

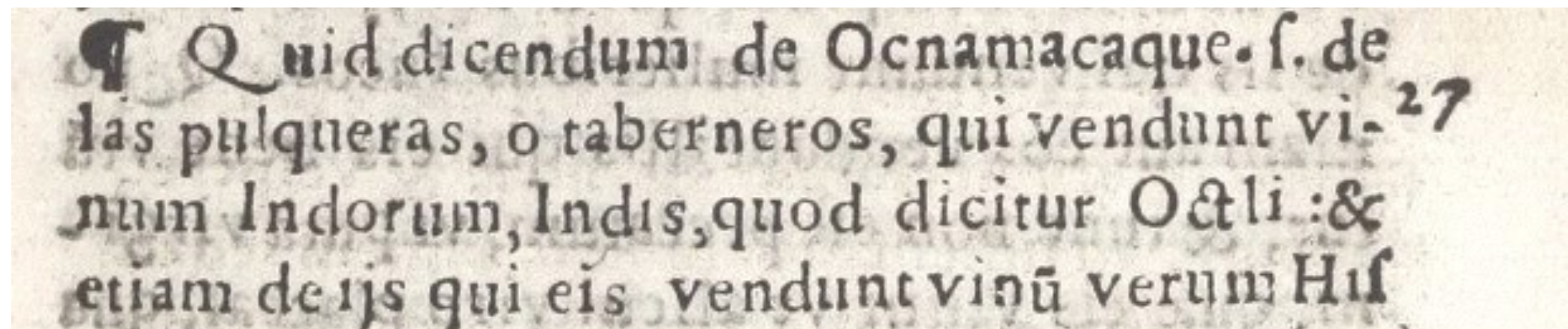
A y proprio vocablo de logro, que es, **tetech-**
tlaixtlapanaliztli, tetechtlamiaccaquixtiliztli,
y para dezir diſte alogro? **Cuix tetech otitlaixt-**
tlapan, cuix tech otitlamiaccaquixti?

Modifying OCR for Colonial Documents

1. Multiple Languages
2. Historical Orthography
3. Modernized Transcription

Multiple Languages

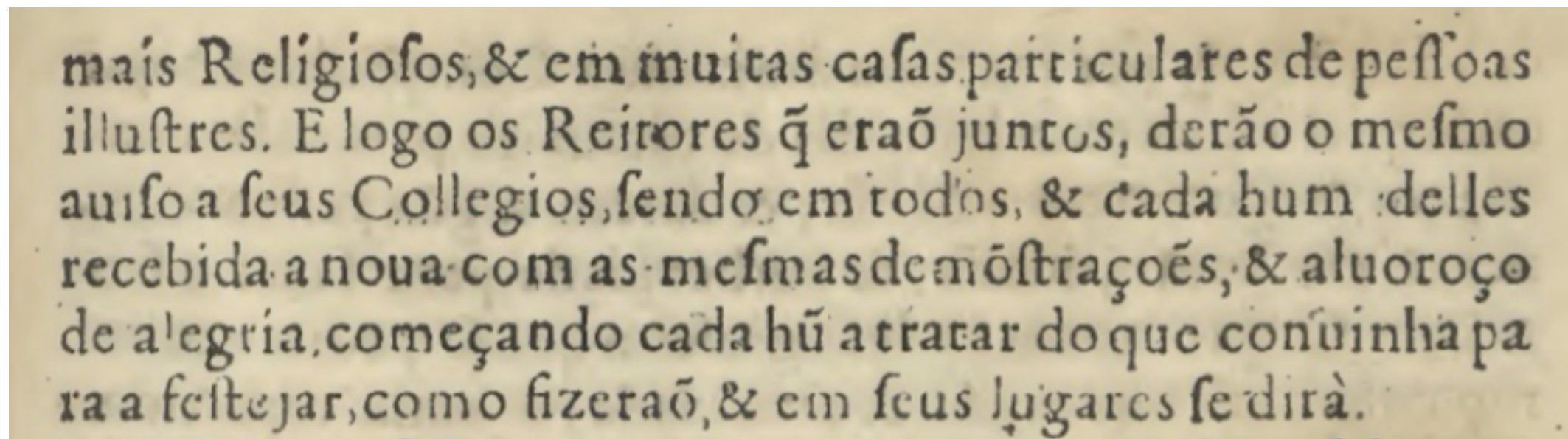
"¶ Quid dicendum de Ocnamacaque. f. de las pulqueras, o taberneros, qui vendunt vinum indorum, indis, quod dicitur Octli .& etiam de iis qui eis venduntvr nō verum Hi



Bautista, *Advertencias*, 1601

Historical Orthography

mais Religio^fos, & em muitas ca^fas particulares de peñoas illustres. E logo os Reitores \~qe taõ juntos, derão o me^fmo avifo a ^feus Collegios, ^fendo em todos, & cada hum delles recebida a nova com as me^fmas demõ^fstrações, & aluoroço de alegria começando cada hũa tratar do que conuinha para a fertejar como fizeraõ, & em seus logares ^fe dirá._



Relações das sumptuosas festas..., Lisbon 1622

[Lisa Voigt, Laura Fernandez-Gonzalez, Iris Kantor, Ângela Barreto Xavier]

Normalization

Norm.: illustres. E logo os Reitores **que** taõ juntos, derão o me**s**mo

Hist.: illustres. E logo os Reitores \~**q**e taõ juntos, derão o me**f**mo

Norm.: avi**s**o a seus Collegios, **s**endo em todos, & cada hum delles

Hist.: avifo a feus Collegios, fendo em todos, & cada hum delles

Norm.: recebida a nova com as me**s**mas dem**on**strações, & al**v**oroço

Hist.: recebida a nova com as me**f**mas dem**õ**ftrações, & aluoroço

Norm.: de alegria começando cada **h**uma tratar do que con**v**inha pa-

Hist.: de alegria começando cada **h**ũa tratar do que con**u**inha pa

illustres. E logo os Reitores q̃ eraõ juntos, derão o mesmo
avifo a feus Collegios, fendo em todos, & cada hum delles
recebida a nova com as mesmas demõstrações, & aluoroço
de alegria, começando cada hũa a tratar do que conuinha pa
ra a festejar, como fizeraõ, & em feus lugares se dirã.

Future Work

Normalization

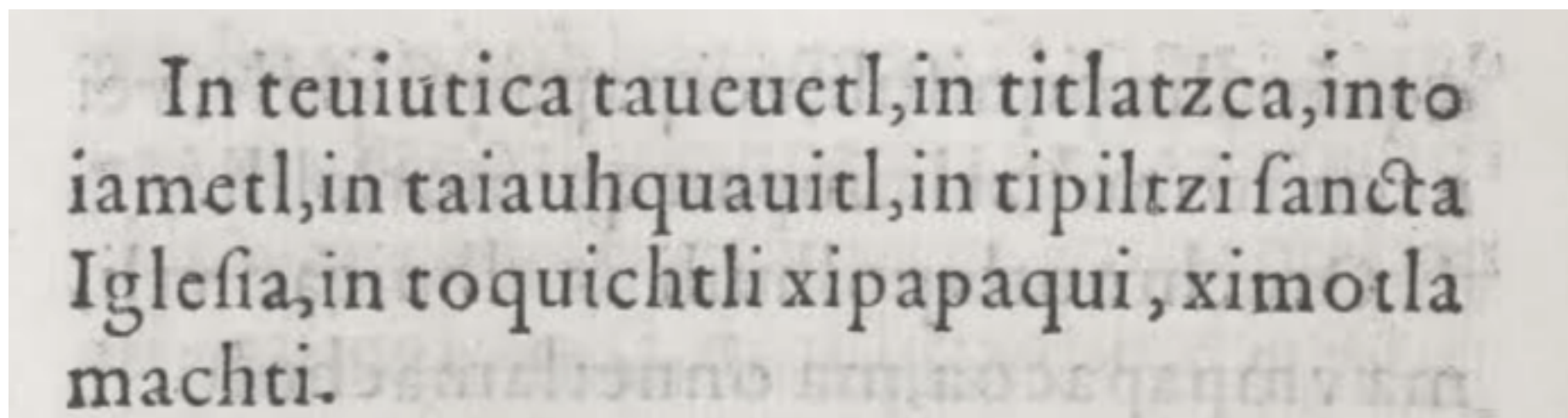
morpheme
parsing

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

teoyotica
+
tawewetl

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

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



In teuiutica taueuetl, in titlatzca, into
iametl, in taiauhquauitl, in tipiltzi sancta
Iglefia, in toquichtli xipapaqui, ximotla
machti.

Sahagún, *Psalmodia*, 1583

Normalization

over -
correction

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

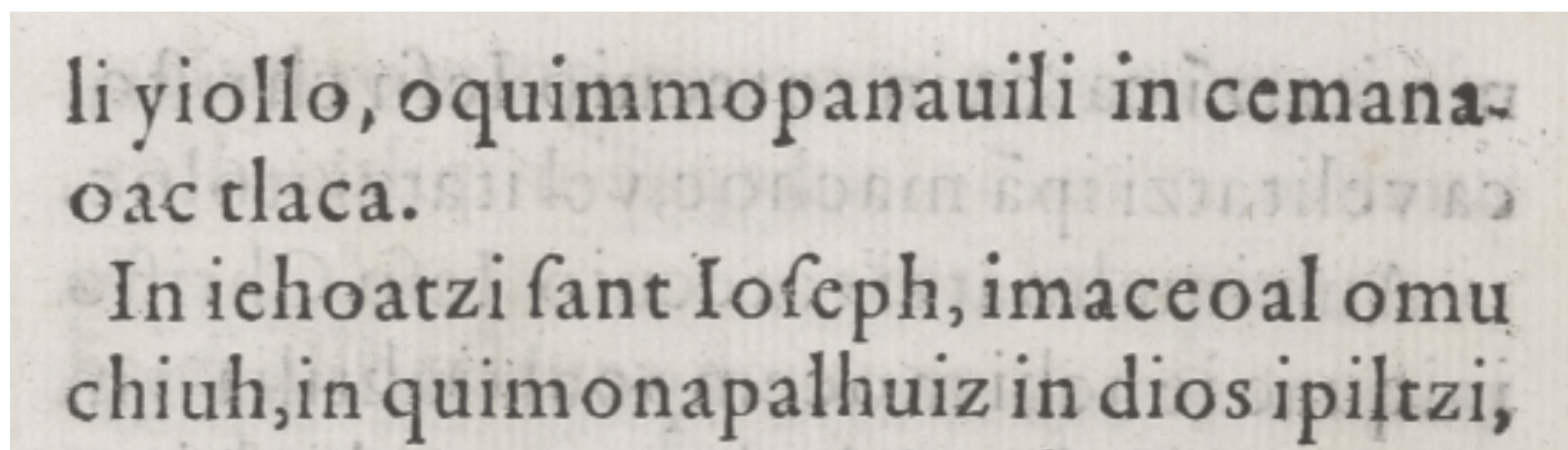
false
substitution

Mod: oac tlaca.
Hist: oac tlaca.

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


correct
substitution!

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



Sahagún, *Psalmodia*, 1583

Integrating with eMOP



IDHMC
 TEXAS A&M UNIVERSITY

FirstBooks Dashboard

Initiative for Digital Humanities, Media, and Culture

Results Filter

Ground Truth All

Works All

Collection All

OCR Batch All

Language All

Print Font All

OCR completed date:
 From: To:

Add Conditions

Reset Filter

Apply Filter

Job Queue

Scheduled:	545
In-Progress:	0
Await Postprocess:	0
Failed:	9
Ingested:	33
Ingest Failed:	0
TOTAL:	587

Select All
 Schedule Selected
 Schedule All
 Set Print Font for Selected Works

Show 25 entries

	Status		Collection	ID	Book ID	GT Number	Language	Title	Author	Font	OCR Date	OCR Engine	OCR Batch	Juxta	RETAS
<input type="checkbox"/>	10-0-0-0		FirstBooks	101	pl_plfx_008		mixtec	Vocabulario en lengua misteca		roman		Ocular	42: 01_testAllBooks	N/A	N/A
<input type="checkbox"/>	10-0-0-0		FirstBooks	102	pl_plfx_009		mixtec	Vocabulario en lengua misteca		roman		Ocular	42: 01_testAllBooks	N/A	N/A
<input type="checkbox"/>	10-0-0-0		FirstBooks	103	pl_tamu_006		mixtec	Doctrina cristiana en lengua misteca		blackletter		Ocular	42: 01_testAllBooks	N/A	N/A
<input type="checkbox"/>	0-0-0-0		FirstBooks	104	pl_tecm_074		mixtec	Vocabulario en lengua misteca		roman				N/A	N/A
<input type="checkbox"/>	10-0-0-0		FirstBooks	105	pl_befk_001		nahuatl	Vocabulario en Lengua Castellana y Mexicana				Ocular	42: 01_testAllBooks	N/A	N/A
								Vocabulario en							

Transcribing the Primeros Libros



“Reading the First Books”

sites.utexas.edu/firstbooks

Hannah Alpert-Abrams
halperta@gmail.com