Automatic Transcription in Colonial Contexts

Hannah Alpert-Abrams, comparative literature
Dan Garrette, computer science
University of Texas at Austin

The *Primeros Libros* project is an effort to digitize all books printed in the Americas before 1601. These books represent early efforts to codify colonial ideology and to produce writing systems for indigenous languages using Latin alphabets and grammatical systems (Mignolo).

Ocular, designed by Taylor Berg-Kirkpatrick et al. (ACL 2013), is the state-of-the-art tool for the automatic transcription of books printed during the hand-press period

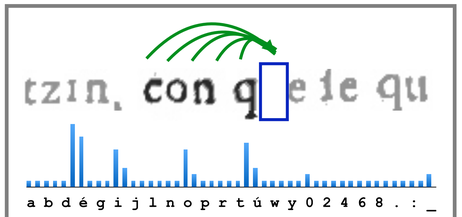
Ocular:
Historical Character Recognition System

Ocular's **font model** takes into account overinking and uneven alignment.

Its **language model** provides context for unclear characters.

Gante (1555)	motlacatilia:ynica sacramento Baptil-
Anunciación (1565)	¶ Picnoneltoquitia ynd̃os
Sahagún (1583)	Yoan oquilhui in Emperador, in tlaca-
Rincón (1595)	ction.v.g.tetlaçotlaliztli.amatio, vel,
Bautista (1600)	limo, hæc supra dictus doctor Medina. Mas

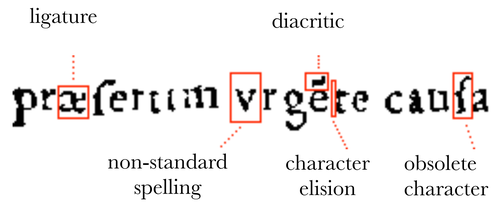
Variation in font, inking, and alignment in the *Primeros Libros collection*



Ocular uses a six-gram language model to analyze context

Extension:
Multilingual Documents Orthographic Variation

Our extension allows the tool to **switch languages** between spaces, and provides an interface for the incorporation of **orthographic variation.**



Examples of common orthographic variations from the *Primeros Libros*.

mentira	mẽtira
mentira	mẽtira

Ocular's language model depends on an external corpus which often "corrects" historical orthographies.

Implications:
Deformation Codification

We argue that the simplicity of our statistical model obscures its engagement with colonial processes through its relationship to deformation and codification.

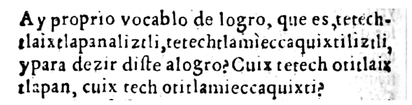
quimomanililique iniuhqui texucutliui
tzo corona:inic macuilpa,iquac in oqui
mococoionilique inimatzi:inic chiqua-
ceppa, iquac in quimococoioniliique
inijcxitzi:inic chicoppa,iquac in quimi-
xiliique yiomotlantzinco.

glutmom an ill liquefiniti liquift executlini exorcoronawinic magnilpa, square in equi- me cocotom liquefinimatzionic gluqua- ge PPa., square in quimoco co fourth lique, in flexitz is interchicop Pa., square in quirmi- xist liquefy somedants inco-

Sample "deformed" output using the *Wall Street Journal* corpus.

Transcription errors alert us to the "deformations" of Nahuatl when written using a Latin writing system

Results:
"Faithful" automatic transcription



A y proprio vocablo de logró, que es . tetech - tlaixtlapanaliztli, tetechtla miec caquixtiliztli, y para decir de|te a logro? Cuix tetech otitlaix- tlapan, cuix tech otitlani teccaquixti?

Sample output from our system, including language switching.

The statistical codification of textual variation validates past orthographies. This rationalization of the historical record reduces the shock of contingency and colonization.

Experimental Results

	Character Error Rate	Word Error Rate
Ocular	12.3	43.6
+ code switch	11.3	41.5
+ orth. var.	10.5	38.2

Lower is better.