Reading the *Primeros Libros*: From Archive to Optical Character Recognition

Hannah Alpert-Abrams
Program in Comparative Literature
The University of Texas at Austin
halperta@gmail.com

Dan Garrette
Department of Computer Science
The University of Texas at Austin
dhg@cs.utexas.edu

Abstract

The PDF images of early American printed books in the Primeros Libros digital collection pose several challenges for Optical Character Recognition (OCR) systems. The Ocular system, designed by Taylor Berg-Kirkpatrick et al., jointly models the physical operation of hand-press printing and the language of the written document, allowing it to learn' to read early printed books. Ocular cannot, however, handle the diacritics and code switching prevalent in the American context. Working with PDF images of trilingual texts in Spanish, Latin, and Nahuatl, we set out to modify Ocular for use on the Primeros Libros collection. Our purpose was, to paraphrase Mary Louise Pratt, to make these texts read (by an online audience) and readable (by a computer).

In this paper, we turn a critical eye to our OCR system. The books in the Primeros Libros collection represent a new print technology that has often been seen as an apparatus of Spanish colonial rule; at the same time, they encode shifting relationships between church and state, between Europe and America, and between Spaniards and indigenous Mexican peoples. In this paper, we will describe how the OCR model transforms these original texts to create a new surface for textual engagement, and how this surface reflects back onto processes of transcription and translation underlying the original production of the Primeros Libros. Focusing specifically on textual code switching and the challenges that it has posed for OCR, We will consider how these technologies engage with processes of isolating and codifying indigenous languages and cultural practices.

Disclaimer: what follows is the written version of a talk presented at the American Comparative Literature Association annual meeting in March, 2015. It is not a published, peer reviewed, or fact-checked article.

1 Introduction

We begin with a page. Figure 1 shows the digital facsimile of a page taken from the *Advertencias para los Confessores de los Naturales*, a confessional manual printed around 1601 in Tlatelolco, Mexico. The image is part of the Primeros Libros collection of digital facsimiles of all books printed before 1601 in the Americas.

A page like this one poses a number of problems for readers of all kinds. It is written in three languages: Spanish, Latin, and the indigenous Mexican language Nahuatl. It is written according to the orthographic conventions of the period, which were far more flexible than today: spelling and spacing are variable, and printers frequently used shorthand to conserve resources. Similarly irregular are its material conditions: smudged, unevenly inked, or misaligned letters, along with the effects of time and the distortions of the scanning process. These factors mean that to read this page requires a certain amount of dexterity on the part of the reader.

The same is true if that reader is a machine.

This paper is concerned with the process of automatically transcribing the Primeros Libros collection through the use of Ocular, a new Optical Character Recognition (OCR) tool developed specifically for use on "historical" printed books by Taylor Berg-Kirkpatrick in 2013. Ocular's innovative approach to OCR, which includes statistically modeling the processes of printing the book as well as the language in which the book is printed, has made it the state-of-the-art for historical OCR. Its statistically simple language model, however, belies the theoretical complexity of mathematically analyzing

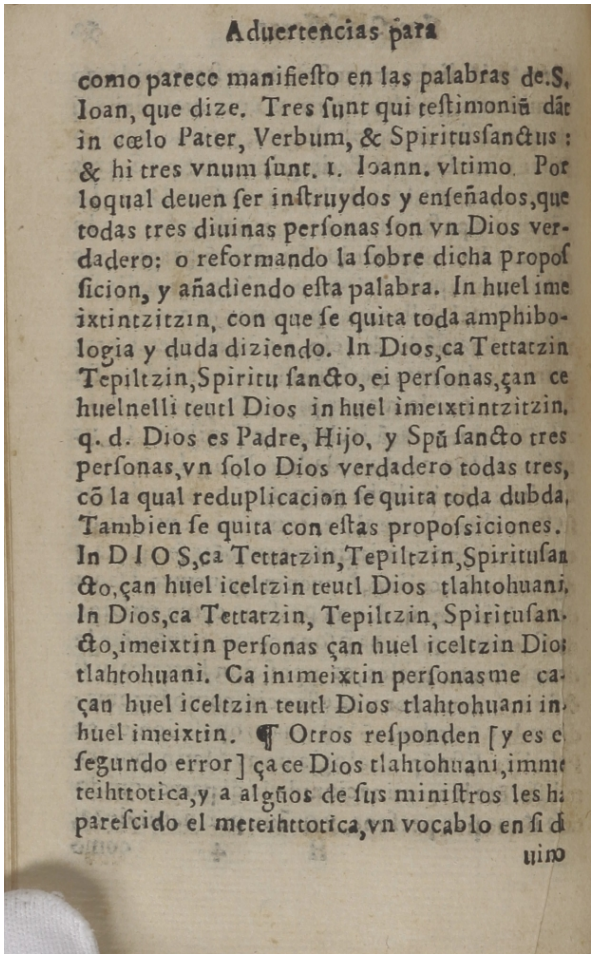

Aduertencias para
como parece manifiesto en las palabras de.S.
Ioan, que dize. Tres ſunt qui teſtimoniũ dãt
in cœlo Pater, Verbum, & Spiritusſanctus:
& hi tres vnum ſunt. 1. Ioann. vltimo. Por
loqual deuen ſer inſtruydos y enſeñados,que
todas tres diuinas perſonas ſon vn Dios ver-
dadero: o reformando la ſobre dicha propoſ
ſicion, y añadiendo eſta palabra. In huel ime
ixtintzitzin, con que ſe quita toda amphibo-
logia y duda diziendo. In Dios,ca Tettatzin
Tepiltzin,Spiritu ſancto, ei perſonas,çan ce
huelnelli teutl Dios in huel imeixtintzitzin,
q. d. Dios es Padre, Hijo, y Spũ ſancto tres
perſonas,vn ſolo Dios verdadero todas tres,
cõ la qual reduplicacion ſe quita toda dubda.
Tambien ſe quita con eſtas propoſsiciones.
In D I O S,ca Tettatzin,Tepiltzin,Spirituſan
cto,çan huel iceltzin teutl Dios tlahtohuani.
In Dios,ca Tettatzin, Tepiltzin, Spirituſan-
cto,imeixtin perſonas çan huel iceltzin Dios
tlahtohuani. Ca inimeixtin perſonasme ca-
çan huel iceltzin teutl Dios tlahtohuani in-
huel imeixtin. ¶ Otros reſponden [y es e
ſegundo error] ça ce Dios tlahtohuani,imme
teihttotica,y a algũos de ſus miniſtros les h
pareſcido el meteihttotica,vn vocablo en ſi d
uino

Figure 1: Folio 52 recto from the *Advertencias para los Confessores de los Naturales*, from an exemplar held by the Benson Latin American Collection.

language in its historical context. As a result, Ocular cannot be easily applied to early modern books that are multilingual or orthographically inconsistent.

In a paper to be presented at the North American Association for Computational Linguistics, we introduce extensions to Ocular's language model that allow the tool to handle multilingual and orthographically complex documents. Here, we consider how these extensions depend on historically specific concepts of distorted, deformed, or corrupted texts. Looking within the context of the Primeros Libros collection, we consider how this concept of deformation plays into a larger history of linguistic codification.

When we speak of deformation, we refer to Jerome McGann's influential work *radiant textuality*. In that work, McGann identified deformation as an analytic tool enhanced by the digitization of aesthetic objects like paintings or works of poetry. He argued that through the process of digitally deforming texts, it becomes possible to perceive new levels of meaning within a work of art. Similarly, we find that the sites of deformation in the automatic transcription process are productive moments that lay bare the underlying complexity of the transcription process.

We focus here on two points of deformation. First, we find that the imperfections of an automatic transcription serve as a deformation of the printed text which draws attention to the relationship between statistical model and historical artifact. At the same time, we see Ocular as a tool which identifies and corrects for certain kinds of distortion in the original text.

This distortion becomes meaningful when it is read against the historical context of the Primeros Libros collection. Written during the early colonial period, the documents in this collection represent some of the first efforts to codify colonial contact using printing technology and alphabetic language. During the sixteenth century, Spanish mendicant friars in the Americas worked to develop alphabetic writing systems for indigenous languages based on the Latin grammatical model. As Walter Mignolo has shown, this project took as a basic assumption the idea that Latin was an ideal language against which all others could be measured; Nahuatl, for example, was found to be deficient in seven letters (46). The shift to alphabetic writing systems also reflects the perceived inferiority of the pictorial writing systems previously in use, a perception which in turn reflects underlying assumptions about the inferiority of indigenous intellectual thought (Boone and Mignolo). We observe here that while the reasons for these assumptions are many and varied, the incompatibility of pictorial writing with European printing technologies is at least reflective of the larger incompatibility between the two systems of communication, and helps to explain the rapid disappearance of the pictorial system from official discourse (though it did continue to be used, in various ways, into the eighteenth century).

Just as the devaluing of pictorial writing carried with it ideological implications, the imposition of alphabetic writing on colonial Mexico reflected and

enacted deeper epistemological conflicts between indigenous and Spanish ways of knowing, as Elizabeth Hill Boone, José Rabasa, Mignolo and others have described. In the work of colonial Nahuatl scholars from Louise Burkhart to Mark Christensen, we see continued attention to the ways that these conflicts are enacted on the written page in Nahuatl or multilingual documents like the *Advertencias para los Confessores de los Naturales* pictured in Figure 1. In this paper, we argue that the shift to alphabetic writing can be understood as a kind of deformation that parallels the deformations of automatic transcription. Though we don't see the work of automatic transcription as a colonizing project,[1] we argue that implicit in these processes are the contexts through which we view the text and the ways in which we intend to read it. Both of these aspects carry ideological weight.

2 Ocular: Historical Document Recognition System

Ocular is an Optical Character Recognition (OCR) tool designed specifically for use on early modern printed books. Most people who work on computers are familiar with OCR: it is used to convert scanned articles in JSTOR, Google Books, and elsewhere into text that can be selected, searched, copied, and underlined. OCR tools work by isolating each character on the printed page and then using a statistical model to identify a correspondence between the printed image and the letters in the alphabet. Though we have all seen the gibberish outputs of inadequate OCR, these tools are relatively successful when applied to contemporary documents for which the correspondence between image and character is fairly straightforward. As Figure 2 illustrates, however, for books printed on a hand-press the correspondence between character and letter is far less easy to model.

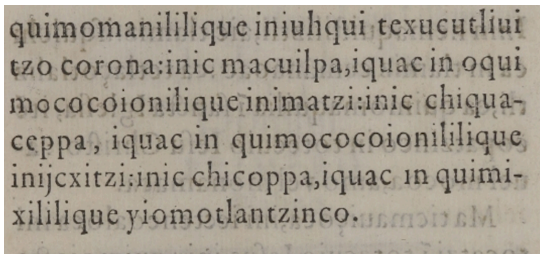
Ocular, which was designed by Taylor Berg-Kirkpatrick et al. in 2013, improves on pre-existing systems for automatic transcription of historical documents in two major ways: the modification of the font model, and the introduction of a language model. The font model is the statistical model that "recognizes" characters by identifying a correspondence between the printed image and characters in a given alphabet; Ocular's innovation is the introduction of a system for handling the distortion caused by over-inked or unevenly aligned type. The language model is a statistical model that uses the context of each character in order to improve accuracy. Though not technically innovative (it is a "simple n-gram model," as one colleague recently remarked), the use of the language model alongside the font model significantly improves accuracy on hard-to-recognize characters.

Figure 2: *Variations of the letter "u" in an early modern book*

In this paper, we bring into question the simplicity of the "simple n-gram model." An n-gram model is a statistical model that determines the likelihood that any particular character will follow a given sequence of characters (of length $n-1$). For example, in its original form Ocular uses a six-gram model based on the *Wall Street Journal*. Given the five-character sequence "winte," Ocular's language model is able to determine that the most likely subsequent character is an "r." In cases where a character in a historical document may be distorted beyond recognition, the language model provides vital information.

The beauty of the n-gram model is that it does not take into account complex concepts like language or grammar. By stripping language of these complicating factors - factors which push up against the concept of "meaning" in uncomfortable ways — the n-gram model implies a neutral approach to language analysis. But as Berg-Kirkpatrick et al. found when they tested their system — and as we found when we sought to apply the system to our corpus of sixteenth century American books — the language model is sensitive to the data on which it builds its statistical model. Since this data is exterior to the corpus,

[1] The relationship between automatic transcription and neoliberal imperialism is complex, and will be addressed in a later work. We acknowledge here the influence of Mara Mills and Wendy Chun in shaping our thinking on this subject.


quimomanililique iniuhqui texucutliui
tzo corona:inic macuilpa,iquac in oqui
mococoionilique inimatzi:inic chiqua-
ceppa, iquac in quimococoionililique
inijcxitzi:inic chicoppa,iquac in quimi-
xililique 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-

Figure 3: Automatic transcription using Ocular with a language model based on the Wall Street Journal corpus. From Bernardino de Sahagún, *Psalmodia*.

it can have a distorting effect on the output of the OCR tool.

2.1 Deformations: Primeros Libros in English

Figure 3 shows an automatic transcription of a page from Bernardino de Sahagún's Nahuatl *Psalmodia* using Ocular's default language model based on the *Wall Street Journal*. The resulting output is a deformation of the original text which looks like gibberish. A closer examination, however, shows that this is not mere gibberish. It is gibberish that looks like Nahuatl, but is linguistically English. The first letter of the Nahuatl text ('q'), for example, has been rendered by the computer as a 'g' because of the visual similarity between the two characters. The sequences of letters, however, draw on the patterns of English language use, producing words that are, if not English itself, at least suggestive of Englishness. The result is a Jabberwocky-like text with a Nahuatl structure.

This multivalent deformation is suggestive of the ways that language ideology more broadly shapes our engagement with early modern texts like those in the Primeros Libros collection. As we described in the introduction, the institution of Nahuatl as an alphabetic language marked an epistemological transformation of the language itself. The result would have been a language that sounded (and signified) like Nahuatl, but looked like Latin: a defamiliarization at the level of the word that parallels the linguistic defamiliarization visualized in the distorted Ocular output. As we face the shock of that distortion, we recognize that it is always present in our engagement with historical documents. We read historical texts through the lens of our own language ideologies, producing deformations of the original text.

2.2 Deformations: Modifying OCR

The deformations of the Primeros Libros collection produced by the *Wall Street Journal* corpus proved theoretically productive, but they are not practical. To automatically transcribe the collection using Ocular, we built a historically appropriate corpus by gathering hand transcriptions of sixteenth-century documents in Spanish, Latin, and Nahuatl. When we tested this corpus on the documents in the collection, however, we continued to see distortions caused by the language model. These distortions were the result of two factors: multilingual documents and orthographic variation.

To address the multilingual nature of the documents, we produced three sixteenth-century corpora (in Latin, Spanish, and Nahuatl) for use on the Primeros Libros collection, drawing on texts from Project Gutenberg and from private collections. We then introduced a system that allows Ocular to select, for any given word, the appropriate language from which to draw its language model. The result is an output that more closely matches historical language usage. The tool also automatically provides each letter with a language tag, making apparent patterns of language use embedded in the text.

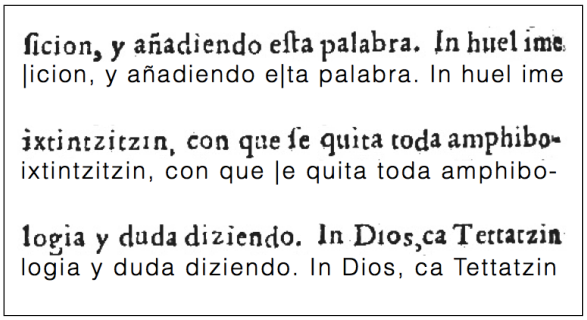
As was described in the introduction to this talk, orthographic variation — spelling, spacing, punctuation, diacritics, and shorthand — is characteristic of all early modern documents, including those in the Primeros Libros collection. When these documents are hand-transcribed, however, they are frequently modified to better match modern conventions. These modifications make the documents easier to use. As a source of statistical data, however, they have a distorting effect on the language model; the result is the automatic "modernization" of the original document. To handle this problem,

we developed an interface for manually modifying the corpus to reintroduce orthographic inconsistency. The interface allows us to distort the corpus, or to undo the corrections wrought by modern transcribers, in order to create "data" that correctly models sixteenth century usage. Though scientifically dubious, this method is effective in improving our transcriptions. At the same time, however, it points to an approach to early modern printing that treats historical language use as distorted or deformed. We might say that this represents a bleeding over of the perceived deformations of the material page (smudged letters, faded pages) into the historically-appropriate irregularities of the words themselves.

The modification of the language model corpus, however, has a corrective effect on this perceived deformation. As the language model more accurately reflects the printed words, thereby producing more accurate transcriptions, we find that it has a secondary effect of making the original orthography appear less "irregular" and more "historically legitimate." This effect gives the appearance of empirically validating the original orthography. We suggest that this points to some of the broader implications of this work: the ways in which statistical models serve to authenticate or mediate our relationship with historical artifacts. Importantly, however, in this case the statistical model itself is not exactly empirical: though the model is correct, the data has been artificially altered to match our external perception of what the output should be.

3 Conclusion

To conclude this talk, we return to Figure 1, which shows a page from the *Advertencias para los Confessores de los Naturales*, one of the books in the Primeros Libros collection. The text is a confessional manual, a guide for missionaries with advice on how to administer the sacrament of confession for indigenous converts. The particular page which has been displayed here — chosen as an example because it is written in Nahuatl, Latin, and Spanish — is a discussion of how to properly communicate the concept of the holy trinity. The risk, of course, is that a new convert might commit unintentional heresy by interpreting the trinity as three separate gods, rather than three facets of a single god.


ſicion, y añadiendo eſta palabra. In huel ime
|icion, y añadiendo e|ta palabra. In huel ime

ixtintzitzin, con que ſe quita toda amphibo-
ixtintzitzin, con que |e quita toda amphibo-

logia y duda diziendo. In Dios, ca Tettatzin
logia y duda diziendo. In Dios, ca Tettatzin

Figure 4: Selection from the *Advertencias* with transcription.

Figure 4 shows a selection of this text, with a transcription beneath it. Loosely translated, the selection reads: "For this reason they must be instructed and taught that all three divine people are one true god, or reforming the abovementioned proposition, and adding this word: In huel imeixtintzitzin, with which all doubt and amphibology [grammatical ambiguity] will be removed."

There are many reasons to dwell on this passage, but for our purposes we are interested in how it draws attention to the distortions of Latinate Nahuatl. The passage argues that the use of a particular Nahuatl phrase can remove all theological ambiguity, but as we know, these Nahuatl phrases carry complex signification that goes beyond what the Christian friars hoped to express. (See the work of Louise Burkhart and others for examples of this complexity). When we look at the transcription, we see a similar effect. The Nahuatl phrase given here reflects a moment of orthographic ambiguity (Alonso de Molina, for example, uses the spelling "uey" rather than "huey") . Like the Nahuatl phrase itself, our transcription promises to remove all ambiguity and doubt: to a general reader, it looks identical to the original. Instead, it preserves ambiguity in the form of textual deformation: in this case, the deformation of the Nahuatl itself. The closer we get to achieving accurate transcriptions, the less obvious the underlying deformations become, and the less explicit the colonial history — and neoliberal present — of our project appear.

Works Cited

Berg-Kirkpatrick, Taylor, Greg Durrett, and Dan Klein. "Unsupervised Transcription of Historical Documents." *Proceedings of ACL*. 2013. Print.

Berg-Kirkpatrick, Taylor and Dan Klein. "Improved Typesetting Models for Historical OCR." *Proceedings of ACL*. 2014. Print.

Boone, Elizabeth Hill. *Stories in Red and Black: Pictorial Histories of the Aztecs and Mixtecs*. Austin: University of Texas Press, 2000. Print.

Boone, Elizabeth Hill and Walter Mignolo, eds. *Writing Without Words: Alternative Literacies in Mesoamerica and the Andes*. Durham: Duke University Press, 1994. Print.

Burkhart, Louise. *The Slippery Earth: Nahua-Christian Moral Dialogue in Sixteenth-Century Mexico*. Tucson: University of Arizona Press, 1989. Print.

Christensen, Mark. *Nahua and Maya Catholicisms: Texts and Religion in Colonial Central Mexico and Yucatan*. Palo Alto: Stanford University Press, 2013. Print.

Garrette, Dan, et al. "Unsupervised Code-Switching for Multilingual Historical Document Transcription." *Proceedings of NAACL*. 2015. Print.

McGann, Jerome. *Radiant Textuality: Literature after the World Wide Web*. New York: Palgrave Macmillan, 2004. Print.

Mignolo, Walter. *The Darker Side of the Renaissance: literacy, territoriality, and colonization*. Ann Arbor: University of Michigan Press, 1995. Print.

Rabasa, José. *Tell Me the Story of How I Conquered You: Elsewhere and Ethnosuicide in the Colonial Mesoamerican World*. Austin: University of Texas Press, 2011. Print.