Unsupervised Code-Switching for Multilingual Historical Document Transcription

Dan Garrette* **Hannah Alpert-Abrams†** **Taylor Berg-Kirkpatrick‡** **Dan Klein‡**

*Department of Computer Science, University of Texas at Austin, dhg@cs.utexas.edu
†Comparative Literature Program, University of Texas at Austin, halperta@gmail.com
‡Computer Science Division, University of California at Berkeley, {tberg,klein}@cs.berkeley.edu

Abstract

Transcribing documents from the printing press era, a challenge in its own right, is more complicated when documents interleave multiple languages—a common feature of 16th century texts. Additionally, many of these documents precede consistent orthographic conventions, making the task even harder. We extend the state-of-the-art historical OCR model of Berg-Kirkpatrick et al. (2013) to handle word-level code-switching between multiple languages. Further, we enable our system to handle spelling variability, including now-obsolete shorthand systems used by printers. Our results show average relative character error reductions of 14% across a variety of historical texts.

1 Introduction

Transcribing documents printed on historical printing presses poses a number of challenges for OCR technology. Berg-Kirkpatrick et al. (2013) presented an unsupervised system, called *Ocular*, that handles the types of noise that are characteristic of pre-20th century documents and uses a fixed monolingual language model to guide learning. While this approach is highly effective on English documents from the 18th and 19th centuries, problems arise when it is applied to older documents that feature code-switching between multiple languages and obsolete orthographic characteristics.

In this work, we address these issues by developing a new language model for *Ocular*. First, to handle multilingual documents, we replace *Ocular's* simple n-gram language model with an unsupervised model of intrasentential code-switching that allows joint transcription and word-level language identification. Second, to handle orthographic variation, we provide an interface that allows individuals familiar with relevant languages to guide the language model with targeted orthographic information. As a result, our system handles inconsistent spelling, punctuation, and diacritic usage, as well as now-obsolete shorthand conventions used by printers.

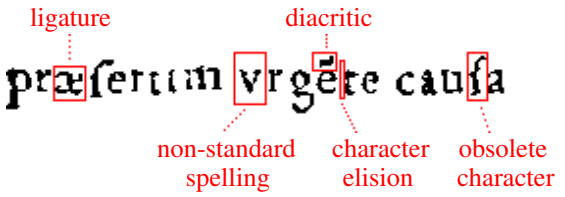
We evaluate our model using documents from the *Primeros Libros* project, a digital archive of books printed in the Americas prior to 1601 (Dolan, 2012). These texts, written in European and indigenous languages, often feature as many as three languages on a single page, with code-switching occurring on the chapter, sentence, and word level. Orthographic variations are pervasive throughout, and are particularly difficult with indigenous languages, for which writing systems were still being developed.

Our results show improvements across a range of documents, yielding an average 14% relative character error reduction over the previous state-of-the-art, with reductions as high as 27% on particular texts.

2 Data

Writing during the early modern period in Europe was characterized by increasing use of vernacular languages alongside Latin, Greek, and Hebrew. In the colonies, this was matched by the development of grammars and alphabetic writing systems for indigenous languages (see Eisenstein (1979) and Mignolo (1995)). In all cases, orthographies were regionally variable and subject to the limited resources of the printing houses; this is particularly true in the Americas, where resources were scarce,

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1036–1041,
Denver, Colorado, May 31 – June 5, 2015.



Input: præſertim vrgẽte cauſa
Language model: praesertim urgente causa

Figure 1: An example OCR input showing the original image and an example of an equivalent modernized text similar to data used to train the LM.

and where indigenous-language orthographies were first being developed (Baddeley and Voeste, 2013).

The 349 digital facsimiles in the *Primeros Libros* collection are characteristic of this trend. Produced during the first century of Spanish colonization, they represent the introduction of printing technology into the Americas, and reflect the (sometimes conflicted) priorities of the nascent colony, from religious orthodoxy to conversion and education.

For our experiments, we focus on multilingual documents in three languages: Spanish, Latin, and Nahuatl. As Berg-Kirkpatrick et al. (2013) show, a language model built on contemporaneous data will perform better than modern data. For this reason, we collected 15–17th century texts from Project Gutenberg,[1] producing Spanish and Latin corpora of more than one million characters each. Due to its relative scarcity, we augmented the Nahuatl corpus with a private collection of transcribed colonial documents.

3 Baseline System

The starting point for our work is the *Ocular* system described by Berg-Kirkpatrick et al. (2013). The fonts used in historical documents are usually unknown and can vary drastically from document to document. *Ocular* deals with this problem by learning the font in an unsupervised fashion – directly from the input historical document. In order to accomplish this, the system uses a specialized generative model that reasons about the main sources of variation and noise in historical printing. These include the shapes of the character glyphs, the horizontal spacing between characters, and the vertical offset of each character from a common baseline. Additionally, since documents exhibit variable inking levels (where individual characters are often faded or smeared with blotched ink) the system also models the amount of ink applied to each type piece.

The generative process operates as follows. First, a sequence of character tokens is generated by a character n-gram language model. Then, bounding boxes for each character token are generated, conditioned on the character type, followed by vertical offsets and inking levels. Finally, the pixels in each bounding box are generated, conditioned on the character types, vertical offsets, and inking levels. In this work, we focus on improving the language model, and leave the rest of the generative process untouched.

4 Language Model

We present a new language model for *Ocular* that is designed to handle issues that are characteristic of older historical documents: code-switching and orthographic variability. We extend the conventional character n-gram language model and its training procedure to deal with each of these problems in turn.

4.1 Code-Switching

Because *Ocular*'s character n-gram language model (LM) is fixed and monolithic, even when it is trained on corpora from multiple languages, it treats all text as a single "language"—a multilingual blur at best. As a result, the system cannot model the fact that different contiguous blocks of text correspond to specific languages and thus follow specific statistical patterns. In order to transcribe documents that feature intrasentential code-switching, we replace *Ocular*'s simple n-gram LM with one that directly models code-switching by representing language segmentation as a latent variable.

Our code-switching LM generates a sequence of pairs (e_i, ℓ_i) where e_i is the current character and ℓ_i is the current language. The sequence of languages ℓ_i specifies the segmentation of generated text into language regions. Our LM is built from several component models: First, for each language type ℓ, we incorporate a standard monolingual character n-gram model trained on data from language ℓ. The com-

[1] http://www.gutenberg.org/

ponent model corresponding to language ℓ is called P_ℓ^{CHAR}. Second, our LM also incorporates a model that governs code-switching between languages. We call this model P^{LANG}. The generative process for our LM works as follows. For the ith character position, we first generate the current language ℓ_i conditioned on the previous character e_{i-1} and the previous language ℓ_{i-1} using P^{LANG}. Then, conditioned on the current language l_i and the previous $n-1$ characters, we generate the current character e_i using $P_{\ell_i}^{\text{CHAR}}$. This means that the probability of pair (e_i,ℓ_i) given its context is computed as:

$$P^{\text{LANG}}(\ell_i \mid e_{i-1}, \ell_{i-1}) \cdot P_{\ell_i}^{\text{CHAR}}(e_i \mid e_{i-1} \ldots e_{i-n+1})$$

We parameterize P^{LANG} in a way that enforces two constraints. First, to ensure that each word is assigned a single language, we only allow language transitions for characters directly following whitespace (a space character or a page margin, unless the character follows a line-end hyphen). Second, to resist overly frequent code-switching (and encourage longer spans), we let a Bernoulli parameter κ specify the probability of choosing to draw a new language at a word boundary (instead of deterministically staying in the same language). By setting κ low, we indicate a strong belief that language switches should be infrequent, while still allowing a switch if sufficient evidence is found in the image.[2] Finally, we parameterize the frequency of each language in the text. Specifically, for each language ℓ, a multinomial parameter θ_ℓ specifies the probability of transitioning to ℓ when you draw a new language. We learn this group of multinomial parameters, θ_ℓ for each language, in an unsupervised fashion, and in doing so, adapt to the proportions of languages found in the particular input document. Thus, using our parameterization, the probability of transitioning from language ℓ to ℓ', given previous character e, is:

$$P^{\text{LANG}}(\ell' \mid e, \ell) = \begin{cases} (1-\kappa) + \kappa \cdot \theta_{\ell'} & \text{if } e = \textit{space} \text{ and } \ell = \ell' \\ \kappa \cdot \theta_{\ell'} & \text{if } e = \textit{space} \text{ and } \ell \neq \ell' \\ 1 & \text{if } e \neq \textit{space} \text{ and } \ell = \ell' \\ 0 & \text{if } e \neq \textit{space} \text{ and } \ell \neq \ell' \end{cases}$$

[2]We use $\kappa = 10^{-6}$ across all experiments.

original	→	replacement
à		a
á		a
que		q̃
per		p̃
ce		ze
x		j
j		x
an		ã
⟨space⟩h		⟨space⟩
be		ve
u		v
v		u
oracion		oron

Table 1: An example subset of the orthographic replacement rules for Spanish.

Finally, because our code-switching LM uses multiple separate language-specific n-gram models P_ℓ^{CHAR}, we are able to maintain a distinct set of valid characters for each language. By restricting each language's model to the set of characters in the corpus for that language, we can push the model away from incompatible languages during transcription if it is confident about certain rare characters, and limit the search space by reducing the number of character combinations considered for any given position. We also include, for all languages, a set of punctuation symbols such as ¶ and § that appear in printed books but not in the LM training data.

4.2 Orthographic Variability

The component monolingual n-gram LMs must be trained on monolingual corpora in their respective languages. However, due to the lack of codified orthographic conventions concerning spelling, diacritic usage, and spacing, compounded by the liberal use of now-obsolete shorthand notations by printers, statistics gleaned from available modern corpora provide a poor representation of the language used in the printed documents. Even 16th century texts on Project Gutenberg tend to be written, for the benefit of the reader, using modern spellings. The disconnect between the orthography of the original documents and modern texts can be seen in Figure 1. To address these issues, we introduced an interface for

author	*Gante*		*Anunciación*		*Sahagún*		*Rincón*		*Bautista*		**Macro Average**		
pub. year	1553		1565		1583		1595		1600				WER
	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER	w.p.
Ocular	13.7	55.9	15.7	53.6	10.8	44.3	11.6	38.4	9.7	25.7	12.3	43.6	56.6
+code-switch	**12.8**	**55.0**	14.6	53.8	9.6	38.7	10.7	35.4	8.8	24.5	11.3	41.5	53.5
+orth. var.	13.5	55.3	**14.1**	**51.6**	**8.4**	**34.9**	**9.5**	**31.0**	**7.1**	**18.2**	**10.5**	**38.2**	**51.0**

Table 2: Experimental results for each book, and average across all books. Columns show Character Error Rate (CER) or Word Error Rate (WER; excluding punctuation). The final column gives the average WER *including* punctuation (*w.p.*). The *Ocular* row is the previous state-of-the-art: Berg-Kirkpatrick et al. (2013). The second row uses our code-switching model, and the third additionally handles orthographic variability.

Gan. (1553)	motlacatilia: ynica ſacramento Baptiſ-
Anu. (1565)	¶ Micnoneltoquitia ynDios
Sah. (1583)	Yoan oquilhui in Emperador, in tlaca
Rin. (1595)	ction.v.g.tetlaçotlaliztli.amatio, vel,
Bau. (1600)	ſimo, hæc ſupra dictus doctor Medina. Mas

Table 3: An example line from each test book.

incorporating orthographic variability into the training procedure for the component LMs.

For our experiments, we built Latin, Nahuatl, and Spanish variability rulebanks by asking language experts to identify spelling anomalies from among several sample pages from *Primeros Libros* documents, and specify *rewrite rules* that map modern spellings back to variant spellings; we also drew on data from paleographic textbooks. Example rules can be seen in Table 1. These rules are used to rewrite corpus text before the LMs are trained; for instance, every nth occurrence of *en* in the Spanish corpus might be rewritten as *ẽ*. This approach reintroduces historically accurate orthographic variability into the LM.

5 Experiments

We compare to *Ocular*, the state of the art for historical OCR.[3] Since *Ocular* only supports monolingual English OCR, we added support for alternative alphabets, including diacritics and ligatures, and trained a single mixed-language model on a combined Spanish/Latin/Nahuatl corpus.

We evaluate our model on five different books from the *Primeros Libros* collection, representing a variety of printers, presses, typefaces, and authors (Table 3). Each book features code-switching between Spanish, Latin, and Nahuatl. For each book, a font was trained on ten (untranscribed) pages using unsupervised learning procedure described by Berg-Kirkpatrick et al. (2013). The font was evaluated on a separate set of ten pages, manually transcribed.[4]

6 Results and Analysis

Our overall results (Table 2) show improvements on every book in our evaluation, achieving as high as 29% relative word-error (WER) reduction.

Replacing *Ocular*'s single mixed-language LM with our unsupervised code-switch model results in immediate improvements. An example of transcription output, including the language-assignments made by the model, can be seen in Figure 2.

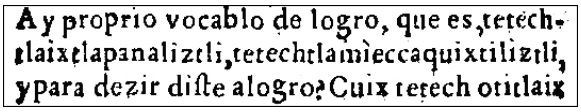
Further improvements are seen by handling orthographic variation. Figure 3 gives an example of how a single spelling variation can lead to a cascade of transcription errors. Here, the baseline system, confused by the elision of the letter *n* in the word *mẽtira* (from *mentira*, "lie"), transcribed it with an entirely different word (*merita*, "merit"). When our handling of alternate spellings is employed, the LM has good statistics for character sequences including the character *ẽ*, and is able to decode the word correctly.

There are several explanations for the differences in results among the five evaluation books. First, the two oldest texts, *Gante* and *Anunciación*, use Gothic fonts that are more difficult to read and feature capital letters that are nearly impossible for the model to recognize (see Table 3). This contributes to the high character error rates for those books.

Second, the word error rate metric is complicated by the inconsistent use of spaces in Nahuatl writ-

[3] http://nlp.cs.berkeley.edu/projects/ocular.shtml

[4] Hyperparameters were set to be consistent with Berg-Kirkpatrick et al. (2013).



Ay proprio vocablo de logró, que es *tetech - tlaixtlapanaliztli, tetechtla miec caquixtiliztli,* y para dezir diſte a *logro? Cuix tetech otitlaix-*

Figure 2: A passage with Spanish/Nahuatl code-switching, and our model's language-coded output. (Spanish in blue; Nahuatl in red/italics.)

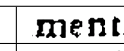
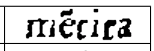
		
no variation handling	mentira	merita
handling variation	mentira	mẽtira

Figure 3: Two variants of the same word (*mentira*), pulled from the same page of text. The form *mentira* appears in the LM training corpus, but the shorthand *mẽtira* does not. Without special handling, the model does not know that *mẽtira* is valid.

ing, falsely claiming "word" errors when all *characters* are correct. Use of spaces is not standardized across the printed books, or across the digitized LM training corpora, and is still in fact a contested issue among modern Nahuatl scholars. While it is important for the transcription process to insert spaces appropriately into the Spanish and Latin text (even when the printer left little, as with *ypara* in Figure 2), it is difficult to assess what it means for a space to be "correctly" inserted into Nahuatl text. *Rincón* and *Bautista* contain relatively less Nahuatl text and are affected less by this problem.

A final source of errors arises when our model "corrects" the original document to match modern conventions, as with diacritics, whose usages were less conventionalized at the time these books were printed. For example, the string *numero* is often transcribed as *número*, the correct modern spelling.

7 Conclusions and Future Work

We have demonstrated an unsupervised OCR model that improves upon Berg-Kirkpatrick et al. (2013)'s state-of-the-art *Ocular* system in order to effectively handle the code-switching and orthographic variability prevalent in historical texts. In addition to transcribing documents, our system also implicitly assigns language labels to words, allowing their usage in downstream tasks. We have also presented a new corpus, with transcriptions, for the evaluation of multilingual historical OCR systems.

Our system, as currently designed, attempts to faithfully transcribe text. However, for the purposes of indexability and searchability of these documents, it may be desirable to also produce *canonicalized* transcriptions, for example collapsing spelling variants to their modern forms. Fortunately, this can be done in our approach by running the variability rewrite rules "backward" as a post-processing step.

Further technical improvements may be made by having the system automatically attempt to bootstrap the identification of spelling variants, a process that could complement our approach through an active learning setup. Additionally, since even our relatively simple unsupervised code-switch language modeling approach yielded improvements to OCR performance, it may be justified to attempt the adaptation of more complex code-switch recognition techniques (Solorio et al., 2014).

The automatic transcription of the *Primeros Libros* collection has significant implications for scholars of the humanities interested in the role that inscription and transmission play in colonial history. For example, there are parallels between the way that the Spanish transformed indigenous languages into Latin-like writing systems (removing "noise" like phonemes that do not exist in Latin), and the way that the OCR tool transforms historical printed documents into unicode (removing "noise" like artifacts of the printing process and physical changes to the pages); in both instances, arguably important information is lost. We present some of these ideas at the American Comparative Literature Association's annual meeting, where we discuss the relationship between sixteenth century indigenous orthography and *Ocular*'s code-switching language models (Alpert-Abrams and Garrette, 2015).

Acknowledgements

We would like to thank Stephanie Wood, Kelly McDonough, Albert Palacios, Adam Coon, and Sergio Romero, as well as Kent Norsworthy for their input, advice, and assistance on this project.

References

Hannah Alpert-Abrams and Dan Garrette. 2015. Reading *Primeros Libros*: From archive to OCR. In *Proceedings of The Annual Meeting of the American Comparative Literature Association*.

Susan Baddeley and Anja Voeste. 2013. *Orthographies in Early Modern Europe*. De Gruyter.

Taylor Berg-Kirkpatrick and Dan Klein. 2014. Improved typesetting models for historical OCR. In *Proceedings of ACL*.

Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein. 2013. Unsupervised transcription of historical documents. In *Proceedings of ACL*.

Thomas G. Dolan. 2012. The Primeros Libros Project. *The Hispanic Outlook in Higher Education*, 22:20–22, March.

Elizabeth L. Eisenstein. 1979. *The printing press as an agent of change*. Cambridge University Press.

Walter Mignolo. 1995. *The Darker Side of the Renaissance*. University of Michigan Press.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Gohneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*.