Automatic Transcription in Colonial Contexts

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

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

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’s font model takes into account overinking and uneven alignment. Its language model provides context for unclear characters.

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

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

Implications:
Deformation
Codification

Examples of common orthographic variations from the Primeros Libros.

Samples “deformed” output using the Wall Street Journal corpus.

Results:
“Faithful” automatic transcription

Sample output from our system, including language switching.

Experimental Results

<table>
<thead>
<tr>
<th>Character Error Rate</th>
<th>Word Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ocular</td>
<td>12.3</td>
</tr>
<tr>
<td>+ code switch</td>
<td>11.3</td>
</tr>
<tr>
<td>+ orth. var.</td>
<td>10.5</td>
</tr>
</tbody>
</table>

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