CATALYST
feminism, theory, technoscience

CRITICAL COMMENTARY
Colonial Copying in an Imperial Age

Hannah Alpert-Abrams
University of Texas at Austin
halperta@gmail.com

In the context of colonial studies, digitization is sometimes hailed as an opportunity for liberation. By digitizing the artifacts of colonial history, we can diversify the cultural record, drawing attention to the ways that contact and conflict are embedded in the everyday practice of textual and material production. Given that many colonial artifacts have multiple cultural affiliations, digitization can also break down the barriers of access that have kept many cultural producers – including those from indigenous, enslaved, and immigrant communities, among others – from engaging with their cultural and intellectual history.

Because digitization is largely replicative and descriptive rather than analytic, we might imagine that digitization workers play a relatively neutral role in the freeing of colonial information. But digitization also always enmeshes cultural objects in a new framework of imperial knowledge production. As we digitize, we call on a sequence of mechanisms for replication, description, dissemination, and display that were designed by and for imperial systems and global capitalist

Alpert-Abrams, H. (2018). Colonial Copying in an Imperial Age. *Catalyst: Feminism, Theory, Technoscience*, *4*(2), 1-6. http://www.catalystjournal.org | ISSN: 2380-3312

corporations. Because the workings of these systems are often opaque, our complicity in these larger systems is difficult to discern.

The colonial Zapotec twitterbot @diidxayooxho, for example, tweets dictionary definitions from the sixteenth-century *Vocabulario en lengua Çapoteca.*[1] In situating these definitions within a social media context, it is making fun of the presentism of social media even as it joins other (human) #UsaTuVoz tweeters in adding indigenous languages to our daily streams. At the same time, the logic that enables @diidxayooxho to serve as a critique of colonialism is dependent on structures created by and for a neo-imperial corporate entity. What does that mean for the creators of @diidxayooxho? For its followers?

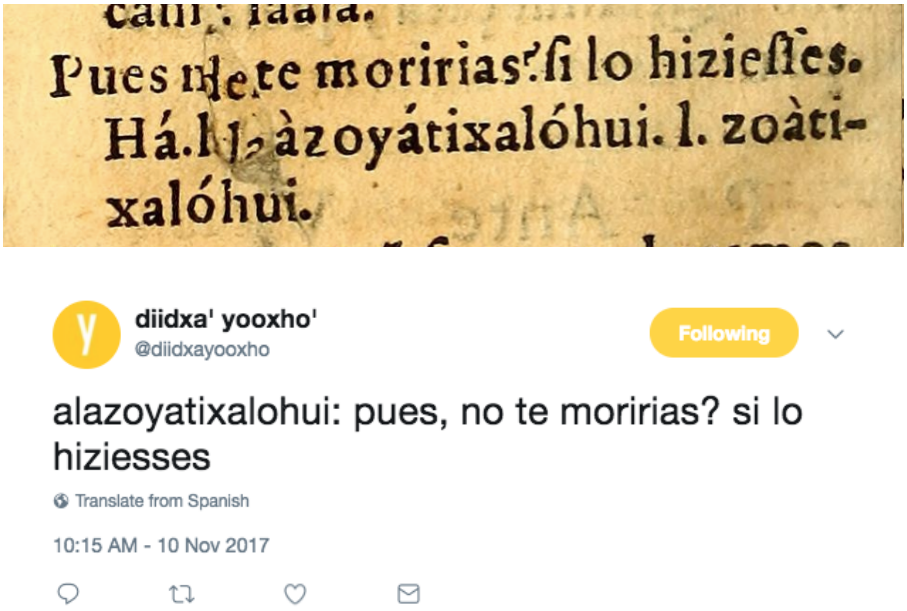

Figure 1. A facsimile of the 1578 imprint of the *Vocabulario*, with associated tweet (Córdova, 1578, p.332 verso; diidza' yooxho', 2017).

Colonial digitization projects invite us to think about our relationship with the ways that discourse has been negotiated in the past and in the present. In the sixteenth century, when Juan de Córdova developed an orthography and a grammar for inscribing Zapotec, a language of southern Mexico, he treated it as a derivative of Latin. Some linguists have interpreted his use of accents as indicating an imperfect

sensitivity to the presence of glottal stops and other non-European features. Smith-Stark (2003, p. 188) argued, however, that the diacritics have no linguistic meaning. He attributed variations in their use to differences in taste between the two compositors who were responsible for setting the type.

Digitization has further distorted these transcriptive features. In the Twitter version, the diacritics have been eliminated entirely to facilitate digital search, and shortened to fit Twitter's rigid character limits ("Acerca del proyecto," 2015).[2] Cultural conflict (Anglophone, Spanish, and Zapotec) and technological contingencies (typesetting, search, and Twitter) shape these transcriptions, both in the colonial dictionary and in its digital instantiation. While these contingencies may appear neutral, they are founded on basic assumptions about language that come from Latin (in the sixteenth century) and English (in the twenty-first).

We see the same mechanisms at work in the case of Optical Character Recognition (OCR), the process by which digital facsimiles are converted into machine-readable text. OCR is one of the many unseen processes that shape a digital edition. It works by combining an analysis of the pixels on the page with a statistical model of language use in order to identify sequences of characters. In the computing world, OCR has been described as a solved problem because the theory that informs it has been proven sound in the case of modern printed documents. OCR is less effective, however, when faced with documents that deviate visually or linguistically from the expectations built into the system. Visual deviation can refer to uncommon typesetting or layout; linguistic deviation refers to character sequences that differ from the statistical norm.

Consider, for example, the following line of text, taken from a document in the *Primeros Libros* collection of books printed prior to 1601 in the Americas. The text begins in Spanish and then switches to Nahuatl, an indigenous language of central Mexico.

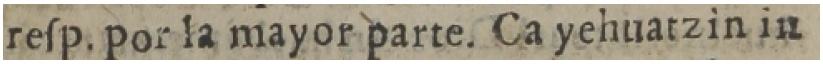

Figure 2. A text in Spanish and Nahuatl.

Now consider three automatic transcriptions of the same line:

1. reſp pool a may or parte. *Ca yeftnat win fit*
2. reſponſortia mayor Paris *Chairman analyfts*
3. reſpi por la mayor parte. *Ca yehuatzin in*

Each of these lines was automatically transcribed using OCR. The first two were created using language models based on English, the system's default language. In the first example, the writings of Lewis Carroll, author of *Alice's Adventures in Wonderland* and *Through the Looking-Glass,* were set as the statistical norm. Here we can see a transcription that is visually Spanish and Nahuatl, but linguistically English: a jabberwocky-like gibberish. In the second example, which uses the *Wall Street Journal* as a statistical norm, the language and places of capitalism are insinuated into the colonial document. The third example, which uses a model I developed in collaboration with computer scientists, is based on historical examples of Spanish and Nahuatl. The misplaced "i" in the first word is caused by the historical use of "resp." in place of the longer Spanish "responden" or even the Latin "respondit". Even with our improved model, historical shorthand remains an unreadable deviation from the norm.

It has been argued that Spanish transliterations of indigenous languages like Juan de Córdova's Zapotec dictionary changed the very shape of indigenous historical memory.[3] The same can be said of automatic transcription, which seeks a compromise between cultural variation and the efficiency of inscription systems. When deviant writing is transcribed with less accuracy than more common linguistic forms, it "dirties" a corpus. This negatively impacts the quality of scholarship on multilingual texts, and places an unequal burden of labor on projects seeking to transcribe deviant texts, like those in indigenous languages. It reinforces the modern, Anglophone status quo for digital projects.

One way to address the colonial implications of automatic transcription is to develop more flexible OCR software. Projects such as *Reading the First Books*, which I managed, are working toward that goal.[4] By collaborating with developers, we hoped to help alleviate some of the ways colonial ideas about language are introduced into digital corpora and to develop tools more sensitive to historical memory. Like the Twitter bot, however, we did so by following procedures and utilizing mechanisms designed to standardize human information for mechanical consumption. Disentangling our work from these procedures and mechanisms, and the beliefs that ground them, is a problem that remains unsolved.

Acknowledgments

Thanks to everyone who had an editorial hand in this project, including the entire decolonial roundtable, especially Mara Mills and Paula Chakravartty; the anonymous reviewers; and the editors at Catalyst. The work based on the *Reading the First Books* project wouldn't have been possible without: Matt Cohen, Kelly McDonough, Kent Norsworthy, Maria Victoria Fernández, Taylor Berg-Kirkpatrick, and Dan Garrette.

Notes

[1] 1 diidxa' yooxho' (@diidxayooxho) is the creation of Rael Albert. See: http://www.iifilologicas.unam.mx/cordova/

[2] "As the website explains: "[Oudijk] observó que todas las etiquetas, marcadores y signos que se habían introducido en las entradas, así como los diacríticos que se habían transcrito del texto original, representaron un problema para las búsquedas. Debido a lo anterior, Oudijk eliminó todos los marcadores, etiquetas y signos…" ("Acerca del proyecto," 2015).

[3] Among many researchers who have considered the history of alphabetization in Mesoamerica are Kathryn Burns, José Rabasa, Camilla Townsend, and Yanna Yannakakis. For a useful survey of recent approaches, see Gabriela Ramos and Yanna Yannakakis' edited volume *Indigenous Intellectuals: Knowledge, Power, and Colonial Culture in*

Mexico and the Andes.

[4] The *Reading the First Books* project is funded in part by a National Endowment for the Humanities Digital Implementation Grant. Any views, findings, conclusions, or recommendations expressed in this article do not necessarily represent those of the National Endowment for the Humanities. You can learn more about the project at sites.utexas.edu/firstbooks.

References

Instituto de Investigaciones Filológicas. (2015). “Acerca del Proyecto.” Retrieved from http://www.iifilologicas.unam.mx/cordova/acerca.php

Córdova, J. (1578). *Vocabulario en lengua çapoteca*. Mexico: Pedro Ocharte and Antonio Ricardo. John Carter Brown Library. archive.org

Ramos, G. and Yannakakis, Y. (Eds.). (2014). *Indigenous Intellectuals: Knowledge, Power, and Colonial Culture in Mexico and the Andes.* Durham: Duke University Press.

Smith-Stark, T. C. (2003). “La ortografía del zapoteco en el vocabulario de fray Juan de Córdova.” In M. de los Ángeles Romero Frizzi (Ed.), *Escritura Zapoteca.* Mexico: Editorial Miguel Ángel Porrúa.

Bio

Hannah Alpert-Abrams is a CLIR Postdoctoral Fellow in Data Curation and Latin American Studies at LLILAS Benson Latin American Studies and Collections, the University of Texas at Austin. www.halperta.com.