HTRC & Me

ACS awardee: Dan Sinykin (Emory University)

In fall 2018, the study of contemporary literature changed. HathiTrust opened its vault of copyrighted material to computational scholarly use. Before then, a scholar wanting to study contemporary literature at scale needed to collect the rare publicly available texts and purchase the rest privately, creating an expensive boutique corpus (text dataset) that could not be shared for collective use or for others to test one's work's replicability: severe limits. Now scholars can access hundreds of thousands of contemporary texts.

Scholars of contemporary literature have—in Matthew Wilkens's oft-repeated phrase—a "problem of abundance." Between five and ten thousand new fiction titles were published each year between 1950 and 2000. The numbers have skyrocketed since then. How can scholars responsibly be accountable to such unwieldy output? There are various good answers; one is to try to account for it all.

I am trying to account for it all by studying the publishing industry. Specifically, I want to know what the conglomeration of the publishing industry did to literature. This would be impossible without computational analysis, and computational analysis would be impossible at the scale I need without the HathiTrust Research Center and the access it gives me to volumes under copyright.

Already, my results have been encouraging. One consequence of conglomeration was the rise of nonprofit publishing. In the late 1970s, literary people worried that the consolidation of publishing into the hands of just a few multinational conglomerates would destroy literature by compelling it to submit to the bottom line. Publishers, the story went, would soon only publish cookbooks, celebrity memoirs, and blockbuster novels—mere entertainments. In response, the National Endowment for the Arts funded small publishers to file as nonprofits. Nonprofit publishers claimed that because they weren't submissive to the bottom line they could publish the literary works that the conglomerates would forgo.

Was it true? Did nonprofits publish fiction that was more literary than that of the conglomerates? Thanks to the HTRC, I could test the claim, and I did. In an earlier phase of my research, I discovered that, yes, nonprofits published fiction that was more literary. (As I have noted elsewhere: "Hathi's holdings are selective, containing those novels acquired by university libraries. I am working with Hathi to assess the histories of transmission that led to its holdings, and my findings are incomplete until I understand how the corporas' omissions shape my results.") But the story is more complicated. Literariness, for the nonprofits, is fiction that foregrounds embodiment. Meanwhile, conglomerates appear to allegorize, in the fiction they publish, the process of conglomeration itself. As I wrote elsewhere, it is as if "the many minds cooperating within modern bureaucracy to bring a book to print composed, beyond their will, a collective agency. Maybe even a collective authorship. Conglomeration expresses itself as mechanical, nonprofits as fleshy. Machine, body. Contemporary literature has a dualism."

But these results are preliminary. The process of taking volumes from the HathiTrust archives and making them available for computational text analysis is difficult. For one, some scholars need to work within a data capsule, accessible through a virtual machine, to keep the copyrighted material legally secure. Working within the capsule poses challenges. Nick Kelly provides helpful accounts here and here.

I managed to build corpora in my data capsule. To do my analysis, I needed to clean the corpora. Here's what that looks like. I held each corpus in a file. One file contains all the novels by Graywolf Press, another by Milkweed Editions, and so on. The Graywolf file contains a file for each novel. Each novel's file contains a .txt file for each *page* of the novel as scanned and OCR'd out of a library. (So, for instance, when I open the file for a 300-page Graywolf novel, I find 300 .txt files.) To use these corpora I needed to do three things: eliminate headers and footers; eliminate front and back matter; and merge page files into a single novel file. The last of these—merging the pages into a novel—is simple to automate. Eliminating front and back matter is too idiosyncratic a task to automate, so it must be done by hand. To eliminate headers and footers, I adapted code from Richard Jean So and Ted Underwood. This code works well, but not as well as it ideally should.

I am working with the HTRC during the 2019-2020 academic year through the HTRC's Advanced Collaborative Support award, which provides staff time for my research project. So far, we have worked together to produce a beta edition of improved code to clean headers and footers from texts. This code will benefit all HTRC users, lowering the bar for doing research with HTRC corpora. Ryan Dubnicek at HTRC has also helped me develop the most comprehensive and up-to-date corpora of HathiTrust holdings of novels based on publisher imprints.

With the new header-footer code and comprehensive corpora in hand, I have begun to replicate my earlier work and extend it to many more publishers. As of now, I have only shown that Random House and four nonprofits can be distinguished by a model. But I am interested in whether this pattern holds, or how it changes, when applied to more representative samples between the two categories. To do this, I will test on corpora from a range of the major publishers from the period, including Harper, Simon and Schuster, Penguin, FSG, and Little, Brown.

I have imported those titles into my capsule, cleaned them of headers and footers, and am currently cleaning them of front and back matter. Soon, I will have a bevy of new results to share. Do conglomerate publishers publish homogeneous fiction? Do imprints have coherent identities? Has conglomeration changed some publishers more than others? Thanks to HathiTrust, I'll soon know.