HathiTrust ACS Grant Write-Up
Richard Jean So

I used my HathiTrust ACS grant (2017-2018) to conduct research towards my current book project, *Redlining Culture: A Data History of American Race and Writing*. This book analyzes the post-war American literary field (1950-2000) through its four major phases: production (book publishing), reception (book reviews), recognition (book sales and book prizes), and consecration (the university). It is mainly interested in tracking how racial inequality – in particular, the extreme whiteness of the literary field – impacts each of these phases, from publishing to reception to canonization, and how that inequality also shapes the literary form, style, and content of this literature (the American novel, in particular). In short, the book seeks to write a new history of race and American fiction after the war by foregrounding the centrality of *inequality*, rather than the ostensible rise of multiculturalism, as it has been traditionally asserted by scholars.

I used the resources of the ACS grant – access to in-copyright digitized library volumes especially – to research and write my first chapter, which focuses on the major American publishing house Random House. Specifically, I was able to compile a comprehensive list of nearly every novel published by Random House from 1950 to 2000, of which the HathiTrust digital library had copies of 1371 texts. I then also compiled metadata on this list regarding the race and gender identity of each author in the corpus, as well as genre information on each of the novels. This data constituted my corpus. With the resources of the Hathitrust's virtual machine environment, which allows researchers to access in-copyright data without consuming or exporting that data outside of the HathiTrust server environment, I was then able to perform automated text analysis – I focused on the use of word embeddings models – on this corpus. I should stress that this corpus is invaluable – it would be impossible to acquire such a corpus (it would likely cost some $30,000 to purchase and/or digitize such texts) on my own and do analysis on it. The HathiTrust's digital library, and its virtual machine environment, which allows researchers to perform computational analysis on it, is enabling unprecedented research.

I did the following analysis: first, I did a simple demographic analysis of the authors. I find, quite surprisingly, that 98% of authors at Random House were white. A mere 1.6% identify or were identified as African-American/Black, and less than 1% identify or were identified as non-Black minority, such as Asian-American or Latino/a. This itself is a major discovery – rarely have we been able to quantify the whiteness of US book publishing. I also discovered that this dominance of white authors is temporally unchanging. With the exception of a small increase in Black authors in the 1970s – overlapping with Toni Morrison's time at Random House as an editor – that increase quickly disappears by the 1990s, and the overall representation of white authors reverts to the mean.

I wanted to know more about the relationship between the demographic dominance of white authors at Random House and the style, content, and form of the novels. I was in particular interested in how this corpus represents racialized characters (white and Black) as a way to understand both the semantic racial tendencies of these writers, as well as how they create worlds of representation and storytelling. I wanted to know: does the

whiteness of Random House have a style and form, and does it tend to engender specific kinds of worlds and characters? First, I wrote a simple algorithm to locate each time a person in a text is racially identified as white or Black, and I plotted these values over time. I found that the overall corpus peaks in racial representation in the 1960s, but that Black authors peak in the 1980s or so. This means that the attention paid to race by white versus Black authors is distinct in this corpus. I also found that Black authors compared to white authors tend to focus on racial representation far more.

Next, I wanted to analyze *how* these characters were portrayed, and here I implemented the word embeddings model, which allows the analyst to identify the words or terms most associated with a keyword or concept of interest. For me, this was the figure of the white character and the Black character in these novels. Somewhat unsurprisingly, I found that the corpus – which again is 98% white – tends to represent Black characters via stereotypes like "thug" and so on, and white characters tend to be described in far more positive terms, like "scientist" or "gentleman." But there were other interesting discoveries. First, the representation of Black characters is highly stable over time. It seems that white authors at Random House are highly limited in their imaginations regarding Black persons in stories, and that also, while sometimes they imagine them in highly negative terms like "thug," often Black characters appear mundane or random. By contrast, white characters are constantly changing and evolving in this period. Most importantly, white characters in this corpus tend to be imagined as intellectuals and writers, and over time, they are increasingly imagined as highly sociable and cultured. In sum: the racial imagination of Random House fiction is intent on having evolving ideas of white characters and persons in their stories, and representing them as literary types (much like the writers who write them) compared to how they think about Black people.

I used these results to write a broader history of Random House fiction and book publishing more broadly in this period, and how it creates specific narratives of race. Here, I combined this computational text analysis with traditional methods in archive, history, and close reading, focusing in particular on Toni Morrison's time as editor. This writing will constitute the first chapter in my book monograph, and will set up my later chapters focused on book reception, book sales, book sales, and the university. In the meantime, I will be presenting shorter versions of this work as public and academic lectures at McGill University and the University of Illinois-Urbana/Champaign this Fall.

I want to reiterate how amazing and valuable this ACS grant opportunity was. Simply, I could not have conducted this research and wrote this book chapter without the resources of the HathiTrust Digital Library and its virtual machine environment. It is truly enabling unprecedented new forms of research on culture and literature. Already I have had many colleagues see or hear about this work and they very much want to get access to this corpus, and other corpora via HathiTrust (in-copyright texts for the post-war period). Upon the release of my book, I will also release the metadata and Hathi code to encourage and enable other researchers to follow up on my work and pursue their own questions. I believe that HathiTrust is facilitating the emergence of an entire field of research.