CA Journal of Cultural Analytics

Race, Writing, and Computation: Racial Difference and the US Novel, 1880-2000

Richard Jean So, Hoyt Long, and Yuancheng Zhu

01.11.19

Peer-Reviewed By: Anon.

Clusters: Race

Article DOI: 10.22148/16.031

Dataverse DOI: 10.7910/DVN/6ANTB8

Journal ISSN: 2371-4549

Cite: Richard Jean So, Hoyt Long, and Yuancheng Zhu, "Race, Writing, and Computation: Racial Difference and the US Novel, 1880-2000," Journal of Cultural Analytics. *January 11, 2019.*

Racial Critique and Data

This article seeks to bridge two scholarly fields often seen as incommensurable: cultural analytics (also known as "computational criticism") and critical race studies.[1] It does so by discovering generative points of contact between two sets

[1] As with previous collaborations, this essay represents the combined efforts of all three authors. While there was some division of labor in the research and drafting of the piece (Richard focused on critical race scholarship and the Bible's place in African-American literature; Hoyt on sociological approaches to race, *Megda*, and black abstractionism; and Yuancheng on refining the statistical modeling), at every stage the labor was informed by collective conversations and experimentation that blurred the presumed boundaries of our respective fields. This perpetual testing of boundaries and knowledge domains represents for us both the excitement and ongoing challenge of this work. We also want to thank our anonymous reviewer, Ronald Judy, Alison Langmead, and all the Novel TM members for their feedback on the project as it evolved.

of methods that are also typically viewed as antithetical: data science and critique. Cultural analytics is an emerging field wherein humanist scholars leverage the increasing availability of large digital corpora and the affordances of new computational tools. This allows them to study, for example, semantic and narratological patterns in the English-language novel at the scale of centuries and across tens-of-thousands of texts. Cultural analytics is a fast-growing field, with scholars taking on an expanding array of topics, including genre and cultural prestige. Yet there is one topic that remains relatively understudied: race and racial difference.[2] The reasons for this elision are not hard to grasp. Computational methods demand the quantification of one's objects of study. It's likely easier to accept measuring a novel's popularity by sales figures or classifying its genre by diction than labeling it according to discrete racial identifiers. Such labeling is an affront to critical race studies, which has taken as its very mission the deconstruction of racial categories.

Unsurprisingly, recent scholarship on the relationship between computation and race has been critique-oriented. Scholars of science and technology, such as Cathy O'Neil and Safiya A. Noble, have documented how computational algorithms used by banks and online search engines intensify racial stratification and oppression by articulating racial minorities as fixed, quantified types that reinforce existing patterns of social inequality.[3] Tara McPherson has shown that the history of modern computation is deeply intertwined with the history of racial formation in the US since the 1960s.[4] Methods born of that earlier computational moment are in part touched by ideologies of racial stratification. This association goes back even further to the birth of social statistics in the nineteenth century, a science whose history is simultaneously the history of efforts to classify people as racial types.[5] One thinks of the intertwining of eugenicist thought with foundational tools for statistical inference, or "general intelligence" tests, both of which were used to assert the inferiority of black people and other racial minorities. Whether reflecting upon the birth of modern social statistics or the current

[2] While the narrower field of cultural analytics has seen relatively little attention to the issues of race and racial difference, our work follows in the path of scholars who have brought much needed attention to these issues as they intersect with the digital humanities more broadly. This includes Elizabeth Dillon, Amy Earhart, Kim Gallon, Jessica Marie Johnson, Angel Nieves, and Roopika Risam, among many others.

[3] Cathy O'Neil, *Weapons of Math Destruction* (New York: Crown, 2016) and Safiya Noble, *Algorithms of Oppression* (New York: NYU Press, 2018).

[4] Tara McPherson, "Why are the Digital Humanities so White?" in *Debates in the Digital Humanities* (Minneapolis, MN: University of Minnesota Press, 2016).

[5] As Tukufu Zuberi notes in his own accounting of this history, "Evolutionary eugenics provided a theoretical context for biological and social statisticians to employ enumerated data for society." See his *Thicker than Blood: How Racial Statistics Lie* (Minneapolis, MN: University of Minnesota Press, 2001), 30.

flowering of machine learning, Albert Murray's warning from 1973 appears as sage advice: "There is little reason why Negroes should not regard contemporary social science theory and technique with anything except the most unrelenting suspicion."[6] Present suspicion bears a historical warrant.

Critical suspicion, of course, can also lead to critical adaptation. As Murray did in his day, much like W.E.B Du Bois before him, and as scholars like Lauren Klein are doing today, the association of quantitative method with structures of racial oppression can be interrogated from the inside.[7] For us, this means understanding the problematic assumptions about race that get encoded in algorithms and computational models and adapting these models in ways that undermine or enrich these assumptions. Inspired by the work of O'Neil, Noble, and social scientists like Du Bois and Murray, we try to imagine what cultural analytics might look like as a method for racial critique. With the increasing visibility of cultural analytics as a field, a commonly heard complaint is: *but what new things does this actually teach us about culture and history?* To which a scholar in critical race studies might add: *what specifically do these methods tell us about the constructed nature of racial difference and identity*? This essay responds to these questions dialectically, using a computational model to study race and literature in order to grasp both the model's affordances *and* its inadequacies. The former provide insight into how race and writing intersect across more than a century of US fiction. The latter are a way to critique and transform the model itself, setting into motion a process of model instantiation and reconstruction that recursively furthers the process of discovery.

This approach is in part inspired by an earlier scholarly attempt to bridge the gap between two fields of humanistic inquiry seen (at the time) as deeply antithetical: critical theory and African-American literary studies. The title of our essay is an homage to the now landmark 1985 special issue of *Critical Inquiry*, "Race, Writing, and Difference," guest edited by Henry Louis Gates Jr. and Kwame Anthony Appiah. In the collection's introduction, Gates outlines the purpose of this special issue as:

> We must, I believe, analyze the ways in which writing relates to race, how attitudes toward racial differences generate and structure literary texts by us *and* about us. We must determine how critical methods can effectively disclose the traces of ethnic differences in litera-

[6] Albert Murray, "White Norms, Black Deviation," *The Death of White Sociology*, ed. Joyce A. Ladner (New York: Random House, 1973), 112.

[7] See especially Klein's "The Image of Absence: Archival Silence, Data Visualization, and James Hemings," which represents an excellent example of this emerging work. In *American Literature* 85, no. 4 (December 2013): 661-688.

> ture. But we must also understand how certain forms of difference and the *languages* we employ to define those supposed differences not only reinforce each other but tend to create and maintain each other.[8]

As Gates writes, the main goal of the issue was to introduce "race" as a keyword into literary studies. Its more specific gambit was to demonstrate the potential utility of new paradigms in critical theory, such as psychoanalysis, for the study of race and literature. "Difference" would be the mediating term. Gates sought to leverage Derrida's deconstruction of speech and writing to unravel otherwise fixed categories of racial identification. However, what most stands out about the issue is the special *care* that its contributors take to stage this encounter. Again and again, they stress that "Third World critics" must critically adopt Western cultural theories, understanding the limits of such theories for non-Western texts, and to use such incongruities to transform both text and theory. To do anything less would be to "substitute one mode of neocolonialism for another."[9] Here we find a useful precedent for our own project.

At the same time, we find in Gates's useful formulation an elision. Of the many new modes of interpretation invoked as the basis for a renewed analysis of race and writing ("Marxist, feminist, post-structuralist"), he avoids mention of methods based in empirical evidence, such as statistics. Most likely, Gates imagined such methods as the simple slotting of things into bins and categories, and thus, diametrically opposed to the project of discerning the contingency of racial categories in order to unravel them. For Gates, quantification meant labeling things (and people), and labeling things meant reifying them. Such a perspective, however, ignores methodological developments in the social sciences which have found sophisticated ways to study race empirically that do not merely produce crude reifications of identity, and which are often informed by the critical race interventions of the 1980s. As humanist scholars interested in adapting quantitative methods to racial critique, we cannot ignore such developments.

Our overall aim in this article, then, is to implement a computational study of race that is critical, reflexive, and interpretative, one that acknowledges the necessary limits of quantitative method (in particular its categorical logic) while exploring its affordances for thinking about racial difference at scale (its patterns and regularities). Indeed, we show how these limits and affordances can mutually inform one another, producing a mode of racial analysis and critique commensurable with foundational critical race studies but also with more recent work in the field by scholars who have sought to destabilize even further the solidity of "blackness"

[8] Henry Louis Gates, "Editor's Introduction," *Critical Inquiry* (Autumn 1985), 15.
[9] Gates, "Editor's Introduction," 15.

as an interpretive category. In concrete terms, we develop a case study focused on race, religion, and the US novel. We build a model to test if novelists marked as "white" versus "black" produce different narratological effects with respect to the interaction of race and religious authority, in particular the authority of the Bible. We then identify a set of general patterns in these effects that we interpret through our model's reliance on reified categories of racial identity. Finally, we propose a method for deforming this very categorical thinking. To wit, we re-read the same large-scale patterns through outliers that trouble this thinking as it exists in our model, but also as it is found lurking by black studies scholars in their field's own interpretive practices. Switching to the kind of relational thinking advocated by some of these scholars, we read categories into continuums, and again into new categories that imagine race as a social assemblage, thereby charting a computational path beyond critique.

Corpus and Model

Our research started with a broad and clear topic—race and the American novel—and a specific question: can computational methods tell us anything new and interesting about how racial difference is expressed in literature? Do authors of different racial identifications (for example, "white" versus "black") consistently use different patterns of language, style, and narrative, and if so, what are these patterns? Do they remain stable or change over time? To begin, we were first drawn to the method known as "sequence alignment." Sequence alignment is usually associated with bioinformatics, where it has been used since the 1980s to track sequences of DNA base pairs with the assistance of computers (Figure 1).

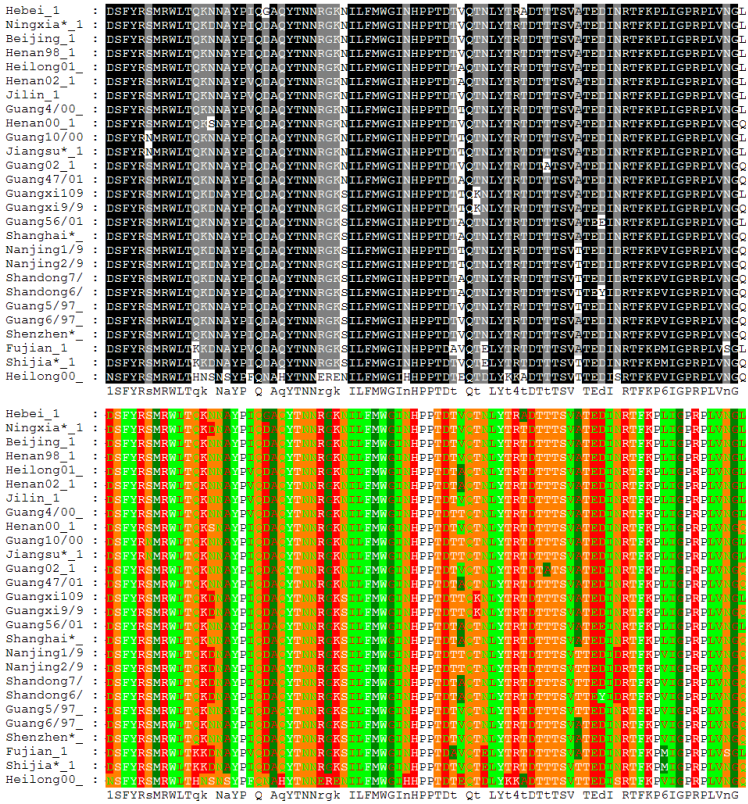

Figure 1. Sequence alignment of avian influenza protein sequences. Algorithms align sequences to identify similar regions that may indicate structural or evolutionary relationships between them. Image source: https://commons.wikimedia.org/w/index.php?curid=965509

But it also has a long history in text search and retrieval, and has been adapted more recently by digital humanists for literary and cultural analysis.[10] Treating objects like novels as long strands of DNA, and words as individual proteins, one can track patterns of language within and across large corpora. One can, for example, find which passages from *Moby Dick* are most cited by American novelists. But oftentimes we want to do more than track the circulation and re-use of a

[10] See, for example, Mark Olsen, et al., "Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections," *Digital Studies* 2, no. 1 (2011); Ryan Cordell, "Reprinting, Circulation, and the Network Author in Antebellum Newspapers," *American Literary History* 27, no. 3 (August, 2015); Kellen Funk and Lincoln Mullen, "A Servile Copy: Text Reuse and Medium Data in American Civil Procedure," in *Legal History* 24 (2016).

single work or author. We want to track the re-use of certain kinds of passages as they are picked up (or not) by certain groups of writers. In our case, we sought to apply the method to the following simplistic questions: do authors identified as "white" versus "black" quote similar types of texts, or even each other? Do they draw from a shared "canon" or radically different sources?

Reviewing the current state of black literary studies, these questions will seem outdated or out-of-step with present scholarly concerns. Scholarship from the mid to late 1980s—largely as an effect of that era's "canon wars"—took an interest in cross-racial literary relations. Hortense Spillers and Robert Stepto wrote at length on the felt literary "kinship" between Ralph Ellison and Ernest Hemingway, while Henry Louis Gates Jr., as part of his major work on black minority discourse, explored the relationship between the Western canon and black literature.[11] But since then, black literary studies have naturally moved on to a new set of questions. Topics such as Afro-pessimism, sound-text dynamics within black literature, black globality, and racial abstraction now command its attention. It's impossible to generalize about the field as a whole, but one might argue that scholars are today less interested in studying literary blackness' relation to whiteness and more focused on probing the autonomous properties of a black aesthetic. We admit that, in broad strokes, the idea of a comparative "white-black" literary analysis might be askance to what scholars in black literary studies now actually care about.

Indeed, such an analysis seems to fall back onto a critical (and political) strategy that underwrote the canon wars: identifying and counting authors as representatives of particular groups. As Gene Jarrett has noted, this strategy "assumed an authentic, if not also autobiographical, relationship between authors and their texts…that what authors *are* predetermines what they will write."[12] We should know better than to try to encode race, which is in reality a highly dynamic and fluid social construction, as a fixed or objective marker ultimately derived from physical and biological characteristics, and ultimately a legacy of the very statistical surveys and methods created to support eugenic thought and racial stratification in the late-nineteenth and early-twentieth centuries. To attempt to count by race presumes there is something there to be counted and is, as demographic historians note, "as much a political act as it is an enumerative one," naturalizing

[11] See Robert Stepto, *From Behind the Veil: A Study of Afro-American Narrative* (Urbana, IL: University of Illinois Press, 2001); Hortense Spillers, *Black, White, and in Color: Essays on American Literature and Culture* (Chicago: University of Chicago Press, 2003); and Henry Louis Gates, *The Signifying Monkey: A Theory of African American Literary Criticism* (Oxford: Oxford University Press, 1988).

[12] Gene Andrew Jarrett, "Addition by Subtraction: Toward a Literary History of Racial Representation," *Legacy* 24, no. 2 (2007): 318.

biological or genetic difference as a fixed, neutral indicator of meaningful social difference between two populations.[13]

Yet as much as this act of counting occludes an understanding of race as ongoing social process, there is some consensus among social scientists that racial categories, empty as they may be, are also strategically useful. Useful because, as participants in the canon wars knew, they represent socially institutionalized essences through and against which social actors must define themselves for the purposes of political mobilization and representation. Race still matters to the extent society remains racially stratified. But social scientists as far back as W.E.B. Du Bois also recognized how useful racial categories are for making quantitative arguments that can combat the effects of racial prejudice by uncovering systemic patterns and trends. They must be critically interpreted, but "to stop collecting racial data prematurely in a racially stratified society would be like *putting the cart before the horse*."[14] Social scientists who work on issues of race thus find themselves in the "treacherous bind" of needing to overturn entrenched racial categories that obscure the dynamics of racial differentiation as lived process while provisionally relying on them as proxies for uncovering these dynamics at larger scales.[15]

Our experiment and data collection all proceeded with these caveats in mind. The novels that form the basis of our experiment are drawn from a larger corpus that was constructed from a list of the most frequently held novels by American authors published between 1880 and 2000 as catalogued by WorldCat. The corpus comprises nearly 10,000 volumes, with peak holdings around 1900 and the 1980s, and represents the work of about 6,000 authors. Given the source, the corpus offers a vision of the literary field based on what librarians have valued, which is in part a proxy for what academics and people who use libraries value. It conveys, as Amy Earhart has argued, only a partial or conditional "truth" about US literature because the "dataset is limited by its construction."[16] The data is biased, of course, and does not provide a neutral account of the American literary field, but at the scale of 10,000 texts, it offers a non-trivial view of this field that

[13] See Melissa Nobles, *Shades of Citizenship: Race and the Census in Modern Politics* (Stanford, CA: Stanford University Press, 2000), 1; and Angela James, "Making Sense of Race and Racial Classification," in *White Logic, White Methods* (2008), 43.

[14] Zuberi, 121.

[15] Yasmin Gunaratnam describes this "treacherous bind" and possible responses to it in *Researching Race and Ethnicity: Methods, Knowledge, and Power* (London: SAGE, 2003), 29-33. Also see Mustafa Emirbayer and Matthew Desmond, *The Racial Order* (Chicago: University of Chicago Press, 2015), who call for a major rebalancing of the theory/method divide in studies of race because empirical methods have long since outpaced theorizing.

[16] Amy Earhart, *Traces of the Old, Uses of the New* (University of Michigan Press, Ann Arbor: 2015), Chapter four. E-edition.

likely captures its dominant trends.

Of the 6,000 authors represented in the corpus, we were able to identify the gender and race of approximately 4,000, which reduced the population of novels on which we could draw since we were only interested in books by authors with marked racial identities. We labeled an author by gender and/or race only if we found that the author self-identified in one particular way and/or if we found such an identification in the scholarly record. Such acts of identification are, most would agree, complicated by the particular exigencies of shifting social and historical circumstance and may bear no direct relation to the novels written under the sign of such identities. But to even begin to interrogate this complexity at scale, it is helpful to start with this crude, provisional form of categorization. It allows us to test whether or not writers were working under essentialized racial categories with any consistency across our period of interest.

After performing this manual labeling, we selected all novels written by authors identified as "black" or "African-American" (this includes biracial authors such as Nella Larsen if we found that African Americanist scholars have identified the author as both biracial and "African-American," as in Larsen's case), which totaled 137 novels for the period 1880-2000. Roughly half of these were published before 1945, and half after 1945. The gender distribution is nearly even, with slightly more male writers. Other categories for which we collected metadata—education, religion, and geography—are randomly distributed. There do exist, however, several biases in the corpus. First, the African-American novel as a genre is known to be diverse in form, and often includes works not traditionally marked as "novelistic." *The Souls of Black Folk*by W.E.B Du Bois is one such example. Our corpus does not include such works because those responsible for getting books into libraries, whose decisions are then reflected in WorldCat, tend to be conservative in what they identify as a "novel." Second, this corpus skews highly canonical. Most of the novels are ones likely to be taught in university classes or included in Norton anthologies. In any case, this corpus represents what we call our "corpus of novels by black authors."

To create a parallel corpus of "white" writers, which far outnumber black writers in the larger corpus, we were careful to select works that similarly skewed canonical. We limited this corpus to canonical writers because otherwise our comparison of "white" and "black" writers would be an anarchic comparison of distinguished black writers against a sea of high and low white writers of all genres. To extract the "canonical" writers, we assembled a list of authors from the *Norton Anthology of American Literature* (2003) and Harold Bloom's *The Western Canon* whom we could identify as "white." About 95% of the authors in this list were in our larger collection, and so their novels (~ 490) became our corpus of

novels by identified white authors. Like our corpus of novels by black authors, the publication dates are split evenly from before and after 1945, but the gender skews more heavily male (about 80% to 20%). Other categories of authorial identity do not skew in any particular direction. In sum, we call these texts our "corpus of novels by white authors." Last, we included the *King James Bible* (KJB) in our corpus after early tests signaled its importance as a shared site of textual quotation.

A final caveat about the data. The choice to preserve an imbalance in our primary classes of interest—there are nearly four times more white authors than black authors—is intended to reflect actual underlying disparities in the history of the modern American novel. Indeed, this 4-1 ratio is likely too conservative. For example, the History of Black Writing project at the University of Kansas, which is constituted by a group of librarians and scholars, has spent the past forty years identifying every novel they could find written by a black person in America from 1880 to 2000. This includes quite obscure novels by authors, for example, who self-published. To date, they have located a total of 1200 novels.[17] By contrast, using far less rigorous methods, we have found more than 6000 novels published by white Americans in this same period—a figure that would no doubt exponentially increase if we had been as exhaustive in our search for white authors as colleagues at Kansas were in their search. If we consider 1200 to be the upper limit for identified black authors, a 4-1 ratio is well beyond conservative.

Next, we elaborate our method: sequence alignment. Genomic scientists use the method to identify DNA patterns across millions of human samples. They might extract, for example, all of the occurrences of the DNA "string" ABCABCABC in a much longer DNA sequence. The same method can be applied to texts as well if we think of sentences as strings of DNA and words or letters as individual nucleotides. Sequence alignment in this case might return every instance of "four score and seven years ago" across millions of documents. Thus it is good at identifying the literal quotation or repetition of lines and phrases, which is an effective way to assess a basic type of textual commonality between white and black writers. While it misses subtler forms of intertextuality like paraphrase and allusion, literal repetition provides a baseline understanding of the degree to which writers of different racial identifications tended to quote the same texts. It identifies a form of commonality based on the direct quotation of a single source.

Several approaches exist for doing sequence alignment, and many have become reliable at detecting near, or "fuzzy," matches. This is important in bioinformatics since genes mutate to produce almost, but not quite, identical sequences. One

[17] Conveyed in private email correspondence with current director of the Project of the History of black Writing, Dr. Maryemma Graham. Spring 2016.

may want a way to find not only complete matches, but something like ABCABD when one was looking for ABCABC. Such "near matches" could represent important cases of genetic mutation. The same goes for texts, which mutate in their own ways. If one is searching "four score and seven years ago" in documents with poor OCR quality, one might also want to find every instance of "pour score and saven years aqo." For this we can use"local alignment" methods that search for smaller matching sequences within much longer ones, often using fuzzy or "close but not quite" matching techniques to do so.

Searching for all of these potential "local alignments" or matches in hundreds of novels is computationally demanding. Fortunately, an efficient approach to sequence alignment has been developed by the ARTFL project at the University of Chicago. It treats documents as ordered sets of *n*-grams ("shingles") by representing each successive, overlapping sequence of *n* words in a document. Trigram shingles are the most common choice, though one can tune this parameter. These shingles are indexed by both document and their position within a document, and are then ordered according to their sequence of appearance. Rather than a sequence of words, a document is represented as a sequence of overlapping trigrams. Importantly, prior to doing so, the program removes high frequency stopwords, short words, and numerals, and can also normalize accents and spelling. This preprocessing eliminates minor textual variations or elisions to allow for more matches.

Once this representation is produced for every document, the program identifies all *exact* three-shingle matches and uses them as starting points for identifying longer sets of overlapping shingles.[18]Upon running sequence alignment on our two sets of novels, the program returned, once duplicates were removed, over 1,200 unique alignments (the algorithm captures bi-directional alignments, thus the need to remove duplicates).[19]To better understand the type of language being shared, or its common source, we labeled all of these alignments and separated them into several categories, including "religious" (any biblical or religious reference; 11%), "lyric" (poetry, religious spiritual, or popular song; 10%), and "self-quotation" (a writer repeating language from an earlier work; 4%). In the lyric category, for instance, an alignment was found between Ralph Ellison's *In-*

[18]For each pair, the algorithm looks at the next shingle in both sequences and checks if they are the same. If they are, it looks at the next one, and so on. When it finds shingles that do not match, it takes note and keeps going until it finds another match. The number of non-matching shingles it looks at before giving up is set by the user and is called the gap parameter. The higher it is, the higher the likelihood of capturing matches across longer sequences. If this gap parameter is exceeded and the resulting alignment is longer than the minimum length requirement set by the user, it is recorded as a local alignment between documents where the sequences occurred. Details on preprocessing options can be found in "Something Borrowed."

[19]All of these are listed in the spreadsheet "ALL_ALIGNMENTS.csv."

visible Man (1952), where he cites the lyric "John Brown's body lies a-mold'ring in the grave, John Brown's body lies…," and a 1931 novel by Booth Tarkington, *The Gentleman from Indiana*, where he cites the very same song: "John Brown's body lies a-mouldering in the ground, John Brown's body lies" (Table 1).

Category of Alignment	Proportion of Total	Example
Religious	11%	Father Which art in Heaven hallowed be thy Name
Lyric	10%	Amazing Grace, how sweet the sound, That saved a wretch
Self-Citation	4%	asked: "Did I snore?" "Terribly," he said, "you sounded like a chain saw
Juridical	4%	find this defendant guilty of murder in the first degree
Quotation	6%	Patrick Henry said 'Give me liberty or give me death'
Aphorism/Saying	2%	to make a long story short
Onomatopoeia	2%	Kitty-kitty-kitty, here kitty-kitty-kitty

Table 1. Example alignments from the over 1,200 unique alignments found across all novels in our corpus. These represent the most common categories of alignment after Bible alignments.

While these results exposed the wide variety of textual patterns shared between novels by white and black authors, an overwhelming number of alignments (550) were direct quotations from the *King James Bible*. If there was a shared language animating literary commonality between white and black novelists in the long twentieth century, it was the language of the Bible.

The Great Code?

To many this finding will not come as a great surprise. Generations of traditional literary and religious studies scholars have painted a heroic picture of the Bible—the King James Version of 1611, in particular—as providing the basis for the Western cultural imaginary. Robert Alter has argued that the "language of the Bible remains an ineluctable framework for verbal culture [in America]."[20] Perhaps most famously, the canonical literary scholar Northrop Frye, as part of his interest in tracking universal literary archetypes, declares that the Old and New Testaments enable "the Great Code of Art"—an expression he borrows from William Blake. Our initial results, in their most naïve and transparent form, appear to confirm what a growing number of traditional scholars of literature and the Bible have argued for some time: that even as the world of the novel grows increasingly sec-

[20] Robert Alter, *Pen of Iron: American Prose and the King James Bible* (Princeton, N.J.: Princeton University Press, 2010), 3.

ular, its commitment to religious ideas and language is never abandoned. It's just altered and transmuted.[21]

Our findings invoke two keywords/concepts—one old, and one new. The first is "universality." The trope of universalism is explicit in Frye's conception of the Great Code. The Bible has a distinct rhetoric, and this rhetoric possesses a "resonance" by which "a particular statement in a particular context acquires a universal significance."[22] Here, an entirely coarse and naïve reading of the results might read as: novels by white and black authors have many obvious and not so obvious differences in matters of form and content, but underlying such differences is a deep structure of universal commonality—the Bible. Both groups of writers mutually quote the Bible. Such a reading invokes a second keyword, a modern inflection of the universalism trope: "virality." Perhaps, one might argue, sequence alignment reveals such strong confluences between different types of texts via the Bible because the Bible has itself a viral quality to it. As Stephen Prickett notes, "grounded in mythologies, belief and texts that have been acquired from elsewhere," the Bible has always required readers and writers to find ways of making these alien words their own.[23] The language of the Bible is "contagious." It has become such a "universal" force in literature because writers find its language so alluring. Both white and black writers find themselves equally stricken by it.

And yet *surely* there must be some differences in how these novelists quote the Bible. Our early attempts to find them, however, came up empty. First, we assessed whether one group cited the Bible more than the other. We randomized the race labels (e.g., white/black) in our dataset and pulled from them a null distribution of quotation counts. We found that the actual amount of Bible quotation by each group was not significantly different from this null distribution. That is, had we assigned the race labels randomly, we could have expected the same rates of quotation. The next dimension we looked at was time. Did white and black writers cite the Bible at different rates over time? Looking at the rates of Bible quotation normalized by the length of the novels, no trends stood out. The data was too sparse for any given year to make any determination. Even when we used a model to predict the amount of quotation given the year and race

[21] Examples of this scholarship include Robert Detweiler, *Breaking the Fall: Religious Readings of Contemporary Fiction* (San Francisco: Harper and Row, 1989); Robert Alter, *Canon and Creativity: Modern Writing and the Authority of Scripture* (New Haven, CT: Yale University Press, 2000); Andrew Tate, *Contemporary Fiction and Christianity* (London: Continuum, 2008); and Amy Hungerford, *Postmodern Belief: American Literature and Religion since 1960* (Princeton: Princeton University Press, 2010).

[22] Northrop Frye, *The Great Code: The Bible and Literature* (New York: Harcourt Brace Jovanovich, 1982), 217-221.

[23] Stephen Prickett, *Origins of Narrative: The Romantic Appropriation of the Bible* (New York: Cambridge University Press, 1996), 35.

of the writer, the resulting curves were inconclusive and biased by a few novels with large amounts of quotation. Within this particular corpus, black writers did not explicitly cite the Bible more or less than white writers at any point in time. Perhaps they were citing different parts of the Bible, which would be interesting given the differences between the ideological orientation of the Old and New Testaments. Yet, when we analyzed whether chapters from either were being cited at different rates, the results were inconclusive.

At this point we began to wonder if the differences did not lie in *what* they were citing, but in *how* they were citing it. We turned our focus to the words surrounding the aligned Bible passages - what we refer to as the alignment "contexts." Using a window of 300 characters on each side of an alignment, we tried to ascertain if white and black writers talked about the Bible differently when they mentioned it. Extracting the contexts for each of the biblical alignments(which gave us ~600 total data points), we treated these as simple word "vectors" and used cosine similarity to measure the lexical difference between every context. We found that the texts did not cluster along racial dimensions, suggesting that white and black writers, as a whole, did not use a different vocabulary when invoking the Bible. Taking this line of investigation further, we fit an LDA topic model on these 600 contexts and represented them as topic distributions across 12 topics. The idea here was that the variation in topics might explain a difference between white and black writers. However, a principal components analysis of the topic distributions showed no distinctive grouping of white versus black authors in the first two principal components.[24] Thus there was no indication that these writers were talking about different things when they cited the Bible. Again, nothing.[25]

[24] The methods we employ here - topic modeling, Principal Components Analysis, and textual classification - are quite common in cultural analytics work. For a good account of topic modeling as used by humanists, see Andrew Goldstone and Ted Underwood, "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us," *New Literary History* (Summer 2014); on Principal Components Analysis, see Paul Vierthaler, "Fiction and History: Polarity and Stylistic Gradience in Late Imperial Chinese Literature," *Cultural Analytics* (May 2016); and on text classification, see Hoyt Long and Richard Jean So, "Literary Pattern Recognition: Modernism between Close Reading and Machine Learning," *Critical Inquiry* (Winter 2016) and "Turbulent Flow: A Computational Model of World Literature," *Modern Language Quarterly* (September 2016).

[25] We recognize that a more robust or sophisticated computational language model might be able to detect other kinds of difference. We provide the full contexts in our Dataverse repository (TAGGED_CONTEXTS.csv) and thus encourage others to pursue alternative methods.

Model Revision I: Against the Great Code

Our first set of results appear to support Frye's "Great Code" thesis: that despite the obvious differences between white and black authors, the Bible joins them to a kind of "common culture." It is a "universal" force of connection, transcending what seem to be nominal differences of racial distinction. In a more contemporary, digitally inflected language, the Bible is a post-racial code, "viral" in its reach. Indeed, the optimism of such a reading is implicitly supported by the felt "universalism" of the computational method that supports such results. Much of the authority and appeal of sequence alignment is that it derives from the methods used to pattern the human genome. As Dorothy Nelkin and Susan Lindeed describe, the language of DNA "pervades our cultural imagination," whereby hyperbolic phrases such as "the code of codes" has imbued the analysis of identity through DNA sampling with profound power and "mystique."[26] Alondra Nelson argues that since the sequencing of the human genome, genetic analysis has acquired a "perceived omnipotence" replete with "seemingly magic powers."[27]

Indeed, the fact that a scientific methodology that espouses "human universalism" so neatly aligns with a fairly conservative argument from literary studies that similarly envisions a sprawling "universalism" to all forms of human expression, should give us pause. Even a cursory reading of scholarship in black literary studies and religion makes clear that black people, due to legacies of slavery and social oppression, hold a complex and embattled relationship to the Bible and Christianity. Theologians such as James Cone point to a central duality: white Christians first introduced the Bible to black subjects in order to enforce docility and acquiescence, an alibi for blacks to reject their concerns for freedom in the human world. But later, black theologians would reinterpret the Bible as mapping a pathway for black liberation, asserting that racial equality represented the will of God, and that its opposite signaled the will of the "anti-Christ."[28]

In terms of literature, black studies scholars will not deny that black writers often cite the Bible or that the Bible occupies a central place in black communities. However, they argue that when black writers invoke the Bible, they do so through a process of "critical modification and revision."[29] This "critical modification"

[26] Cited in Alondra Nelson, *The Social Life of DNA: Race, Reparations and Reconciliation after the Genome* (Boston: Beacon Press, 2016), 4.

[27] Nelson, *The Social Life of DNA*, 4.

[28] James H. Cone, *Black Theory and Black Power* (Ossining, NY: Orbis Books, 2008), 120.

[29] Tuire Valkeakari, *Religious Idiom and the African American Novel, 1952-1998* (Gainsville, FL: University of Florida Press, 2007), 29. See also James Coleman, *Treatments of the Sacred, Spiritual, and Supernatural in 20th-Century African American Fiction*(Baton Rouge, LA: Louisiana State University Press, 2009), 81.

takes several forms: (1) irony: when the Bible is evoked, its meaning becomes ironic because one sees that its normative meaning does not ideally suit the specific conditions of a black character or context. (2) Criticism: when the Bible is cited or invoked, its meaning is explicitly criticized by a black character, revealing some innate hypocrisy or contradiction in the text's message when applied in an African-American context. (3) Dialogism: when the Bible is cited or invoked, it is not delivered as a monologic polemic or sermon, but rather, its invocation immediately incites debate or dialogue over its meaning by a cohort of characters, including potentially the narrator him or herself. In sum, scholars argue that the appearance of the Bible and its quotation in novels by black authors tend to be very dialogic and interactive, whether that is the interaction between characters in the story or the interaction between the reader and the story's characters, all in an effort to question the Bible's normative or hegemonic "white" meaning.

Our attempts to identify differences in the language used around quotations of the Bible had failed to find at least one important signal in the data. We decided to revise our approach to capture this sense of dialogism or "sociality" in Biblical quotation—i.e. *how* the Bible is cited. Specifically, we used content analysis to revisit the "contexts" we had extracted for each biblical alignment. In each context, we looked for moments of sociality, here defined as the presence of two or more characters engaged in dialogue or interaction. If we could find such an instance, we coded the alignment as 1. If we could not find such an instance, we coded the alignment 0, meaning that there was no indication of sociality. In some cases, this meant a lone character delivering a monologue, while in others it was a character lost in thought. For each context, we now had a variable called"social," coded either 1 or 0. As short-hand, we refer to the Bible quotation and its surrounding context as a "scene." Our "social" variable captures whether and *how* white versus black writers infuse that scene (the moment when the Bible appears) with "sociality" (Table 2).

Bible Context	Label	Original Novel	Author
While Davis chanted a traditional prayer-poem with his own variations, Joe mounted the box that had been placed for the purpose and opened the brazen door of the lamp. As the word Amen was said, he touched the lighted match to the wick, and Mrs. Bogle's alto burst out in: We'll ##walk in de light, de beautiful light Come where the dew drops of mercy shine bright Shine all around us by day and by night Jesus, the light of the world##. They, all of them, all of the people took it up and sung it over and over until it was wrung dry, and no further innovations of tone and tempo were conceivable. Then they hushed and ate barbecue.	Social	Their Eyes were Watching God	Zora Neale Hurston
When it was all over that night in bed Jody asked Janie, "Well, honey, how yuh it?" "No," said Adam, "you're making fun." "I'm not," said Miranda, "I'm trying to keep from going to sleep. I'm afraid to go to sleep, I may not wake up. Don't let me go to sleep, Adam. Do you know Matthew, Mark, Luke and John? Bless the bed I lie upon?" "If I should ##die before I wake, I pray the Lord my	Social	Pale Horse, Pale Rider	Katherine Anne Porter

Bible Context	Label	Original Novel	Author
soul to take##. Is that it?" asked Adam. "It doesn't sound right, somehow." "Light me a cigarette, please, and move over and sit near the window. We keep forgetting about fresh air. You must have it." He lighted the cigarette and held it to her lips. She took it between her fingers and dropped it under			
of course. But in the end there returned the poignant yearning from the Sunday world. As she went down in the morning from Cossethay and saw Ilkeston smoking blue and tender upon its hill, then her heart surged with far-off words: "Oh, Jerusalem, Jerusalem-how ##often would I have gathered thy children together as a hen gathereth her chickens under her wings##, and ye would not-" The passion rose in her for Christ, for the gathering under the wings of security and warmth. But how did it apply to the weekday world? What could it mean, but that Christ should clasp her to his breast,	NonSocial	The Rainbow	D.H. Lawrence
as a mother clasps her child? And oh, for Christ, for him safe, still free, still searching-when she awoke. Awoke where she sat now. She ran her fingers through her hair, then stretched and tried to remember why she was here. Her Bible lay on the couch before her, and a ray of early morning sun shone through the window onto the page. "##Let not your heart be troubled: ye believe in God, believe also in me. In my Father's house are many mansions: if it were not so, I would have told you. I go to prepare a place for you. And if I go and prepare a place for you, I will come again, and receive you unto myself; that where I am, there ye may be also. And whither## I go ye know, and the way ye know." Then her eyes moved down the page: "If you love me, keep my commandments. And I will pray the Father, and he shall give you another Comforter, that he may abide with you for ever; even the Spirit of truth; whom the world cannot receive, because	NonSocial	Passing by Samaria	Sharon Ewell Foster

Table 2. Four example contexts where the Bible is cited (indicated by ##). Two are coded as "social" to indicate a scene of dialogic interaction or the presence of multiple characters. Two are coded as "non-social" to indicate a scene where the Bible is being read by a single individual or cited as part of an interior monologue.

The next step was to build a statistical model to see if we could predict if a scene was "social" or not based on several variables. For this task we used logistic regression, which is a standard statistical approach to analyzing the relationship between variables. Such a model determines the likelihood that an object belongs to one category or another based on the predictor variables provided, and in doing so finds which variables are significant or useful in making that prediction. Each object will receive a score between 0 and 1: if the score is above 0.5, the object belongs to the first class, and vice versa. For example, one can use such a model to predict whether a flower is of one type or another (0 or 1) based on a few predictor variables, such as petal length and width.

For our model, we had three predictor variables: "gender" (the author's gender); "race" (the author's race); and "Bible," which indicates whether a scene has a quotation to the Bible or not. The "Bible" variable acts as a control. We wanted to be sure that if we saw an effect around gender or race in passages citing the Bible that this effect was tied to the fact that the Bible was being cited, and not that novels by

white or black authors are inherently more "social." Thus, we randomly selected 400 passages from novels by each group of writers, each the same length as our alignment contexts and split evenly between the two groups. We tagged them by race, gender, and "social," the same as we did with our 600 Bible quotation passages, giving a total of 1000 contexts. The "Bible" variable indicates whether the Bible is cited or not. With this control added, if we found that the identified race of an author helped predict if a Bible scene was "social" or not, we would know it was because that scene had the Bible in it and not because novels by black writers simply have more "social" scenes. Last, we worried that single novels might be contributing a disproportionate amount of Bible quotations, so we added a random effects variable, insuring that no single novel became the source of any specific effect.

$$Y_i \sim \text{Bernoulli}(p_i)$$

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 \cdot \mathsf{gender}_i + \beta_2 \cdot \mathsf{race}_i + \beta_3 \cdot \mathsf{bible}_i + \beta_{23} \cdot \mathsf{race}_i : \mathsf{bible}_i + \gamma_{\mathsf{novel}_i}$$

Figure 2a. The logistic regression model we used to understand the relationship of race, gender, and biblical citation to the "sociality" of contexts where the Bible is quoted.

		Estimate	Std. Error	z score	p value
$\hat{\beta}_0$	(intercept)	0.520	0.265	1.966	0.0494
$\hat{\beta}_1$	(gender)	−0.313	0.237	−1.322	0.1863
$\hat{\beta}_2$	(race)	−0.220	0.252	−0.871	0.3837
$\hat{\beta}_3$	(bible)	−1.328	0.222	−5.994	2.05e−09
$\hat{\beta}_{23}$	(race : bible)	1.742	0.355	4.904	9.41e−07

Figure 2b. Estimates produced from the above model.

Here is the specified model and the estimated coefficients for each term in the model (Figure 2).[30] In interpreting these results, we are interested in the interaction between the "Bible" and "race" variables. That is, we are interested in whether the odds of a Bible context being "social" increases or decreases based on the race variable: whether it is 1 or 0, or "white" or "black." In plain language, we want to know if the fact that a writer is identified as white or black, when that writer cites the Bible, significantly changes the likelihood that the scene of citation is "social." Or, in plainest language: do writers so identified contextualize the Bible in different ways in the novel? Our results suggest they do. When a white writer quotes the Bible, according to our results, it is less likely that she/he quotes it in a social context. In fact, the odds of being "social" decreases by a factor of 3.8,

[30] A fuller description of the model and our procedure for calculating the odds ratios can be found in "Appendix1.pdf" in our Dataverse repository.

compared to when she/he writes about non-Bible related topics. When a black writer quotes the Bible, it is more likely that she/he quotes it in a social context. The odds of being "social" increases by a factor of 1.5 compared to when she/he writes about non-Bible related topics. In sum, white and black writers tend to narrate the Bible differently when they cite it.

This new set of results usefully nuance our first findings. Our first set of results make a valuable discovery: the Bible represents *the* major site of shared discourse between white and black authors in the twentieth century. It is not, as previous scholars have argued, canonical authors like Shakespeare or popular music.[31] Further, this form of discourse does not have clear forms of differentiation. For example, white or black writers, at least those represented in our corpus, are not more likely to cite the Old or New Testament, or any specific Biblical book. However, our second approach exposes a core oversight in the first, namely its inattention to the context of quotation. Taking a cue from existing scholarship, we confirm with our second approach that the forms of sociality attached to each quotation challenge Frye's notion of "the Great Code." There is a rupture in this broader pattern - a pattern within the pattern - that marks a distinction in how white and black writers narrate the Bible. If this pattern was already intuited by black studies scholars, our results extend that intuition to the scale of hundreds of novels written over a century and suggest it can be captured by a simple variable: "sociality."

Model Revision II: Against Reification

Thus far, our proposed computational approach to studying race and the US novel has served to reveal a set of regularities at scale. This revelation, however, has come at the expense of slotting novels into discrete categories: "novels by white authors" and "novels by black authors." For our model to work, writers *must be* white or not white, black or not black. We then find that writers assigned to one category share certain commonalities and differences with writers assigned to the other. Lacking any deeper analysis or critique of these assigned categories, our method could become circular if we presumed the reality of racial categories and their correlation with particular literary effects, as if the categories themselves were causing these effects.

[31] See Gates and more recently, T. Austin Graham, *The Great American Songbooks: Musical Texts, Modernism, and the Value of Popular Culture* (New York: Oxford University Press, 2013).

As race scholars in the social sciences have pointed out, this has long been the problem when racial categories are deployed in quantitative studies. The race variable is interpreted in the absence of a nuanced underlying theory to explain the mechanism by which race effects social outcomes.[32] Filling in for this absence is a generic notion of race rooted in assumptions about biological or genetic differences between populations. When these are treated, by default, as the causal mechanism behind a set of relationships, the very dynamism of race as a social construct is obscured.[33] The result is a study that, for instance, finds an association between "black" populations and heart disease and interprets race as the cause of the disease. In reality, what may be contributing to the association are a host of other variables that interact with race in a particular context, including socioeconomic status, cultural factors, and levels of access and accessibility to health care.[34] To avoid reinforcing the notion of race as a fixed characteristic, these same critics propose various ways to triangulate between quantitative and qualitative methods, leveraging the temporary closure provided by racial categories precisely to open them up to more situated or elaborate accounts.[35]

[32] Zuberi writes that, "Statistical results, themselves, do not prove anything beyond the numerical relationship between two or more lists of numbers or variables" and that how we understand "the connection of these variables in the real world" always requires an underlying theory. Unfortunately, that theory often goes unarticulated when interpreting race as a variable. See "Toward a Definition of White Logic and White Methods," 9. Angela James puts it most bluntly when she writes that race has become a "black hole" of social scientific research and argues that "the use of race as a control variable flattens out the meanings of racial differences and replaces it with a generic notion of difference." See "Making Sense of Race and Racial Classification," 43.

[33] As Zuberi puts it, "Race is not about an individual's skin color. Race is about an individual's relationship to other people within society." Zuberi, "Toward a Definition of White Logic and White Methods," 7.

[34] Wilkinson and King, "Conceptual and Methodological Issues in the Use of Race as a Variable: Policy Implications," 60. On the problem of measuring human variation in a post-genome era, also see Fatimah Jackson, "Anthropological Measurement: The Mismeasure of African Americans." According to Zuberi, this does not mean race cannot be used as a variable, but that "statistical models that present race as a cause are really statements of association between the racial classification and a predictor or explanatory variable across individuals in a population." Association is evidence of causation only when "it is buttressed with other knowledge and supporting evidence" of the many "contingencies or circumstances" which can influence social outcomes. Zuberi, *Thicker than Blood*, 129, 133. For Zuberi, the ability to interpret race as a causal mechanism requires that it be viewed as a "treatment" (i.e., something that can be manipulated) which can be given or denied to the populations under study. But as Paul Holland observes elsewhere, "Properties or attributes of units are not the types of variables that lend themselves to plausible statements of counterfactuality. For example, because I am a White person, it would be close to ridiculous to ask what would have happened to me had I been Black." See "Causation and Race," 100.

[35] John Stanfield argues for a triangulation of quantitative and qualitative methods, as "the former are important for trend analyses and for formulizing patterns and exceptions, particularly when using large data sets; the latter are important for capturing deeply rooted immeasurable subjective experiences such as emotions and spirituality, so crucial for grasping racialized experiences." See *Rethinking Race and Ethnicity in Research Methods* (Walnut Creek, CA: Left Coast Press, 2011), 23. Similarly, Yas-

Something of this critical thrust is present in Alexander Weheliye's revisionist take on "black studies" conceived as a mode of knowledge production. In *Habeas Viscus*, he asserts, "Continuing to identify blackness as one of black studies's primary objects of knowledge with black people as real subjects… accepts too easily that race is a given natural and/or cultural phenomenon." We should approach race instead as an assemblage of forces that "must continuously articulate nonwhite subjects as not-quite human."[36] From a social scientific perspective, one way to think about such assemblages would be to develop more complex quantitative models. That is, we could replace race as a single variable with a host of other variables that we believe might capture the interaction of racial experience with religious practice and literary form. As literary critics, however, we recognize that this interaction is not like heart disease. It is not so easy to identify and collect reliable data on the variables (e.g., religious affiliation, socioeconomic status, spiritual influences) that potentially inform an author's Biblical disposition and which are further mediated by the process of fictionalization itself.

How then to de-reify our interpretation of race while not wholly disregarding what we learn from its temporary closure? Here we propose using information contained in our model, reductive as it is, such that we can pivot from a "grammar of comparison," to use Weheliye's terms, to a grammar of relationality.[37] We leverage the numerical association between race and Biblical citation to *relate* our texts in ways that complicate the grammar of categorical racial difference initially imposed on them. With this relational framework, we are able to read the interaction of race and biblical citation across the color lines that, in both quantitative and qualitative approaches, have tended to segment and particularize the ontological totality of human social relations.

The way we shift the grammar of analysis is through a measure latent in our statistical model. It is a coefficient derived from our regression formula that indicates the relative "sociality" of a text. The measure captures how well an individual text conforms to the associations between race, gender, and biblical citation that

min Gunaratnam insists that, "while their must be temporary moments of closure in the defining of racial and ethnic categories in order to do research, these points of closure must also…be opened up in ways that enable us to look at and hear *how* 'race' and ethnicity are given situated meaning within accounts." Gunaratnam, 38. Zuberi integrates this process of opening up within the statistical model itself, though to do so "requires an elaborate theory that states explicitly and in detail the variables in the system, how these variables are causally interrelated, the functional form of their relationships, and the statistical quality and traits of the error terms." Zuberi, "Deracializing Social Statistics," 131.

[36] Alexander Weheliye, *Habeas viscus: Racializing Assemblages, Biopolitics, and Black Feminist Theories of the Human* (Durham, NC: Duke University Press, 2014), 18-19.

[37] Weheliye, *Habeas viscus*, 13. Weheliye argues that to theorize racially motivated acts of political violence through comparison, rather than in relational terms, reaffirms existing hierarchies rather than realizing articulations of an ontological totality composed through social relations.

we discovered in the overall corpus. For instance, a text by a black writer that frequently quotes the Bible and only in a "social" way will score high on this measure. Conversely, a text by a white writer who quotes the Bible frequently in a non-social way will score very low. These scores allow us to pivot between individual works and the background trends evident in the data, but also to relate them to each other in ways that loosen the link between text and race as a categorical label (Figure 3).

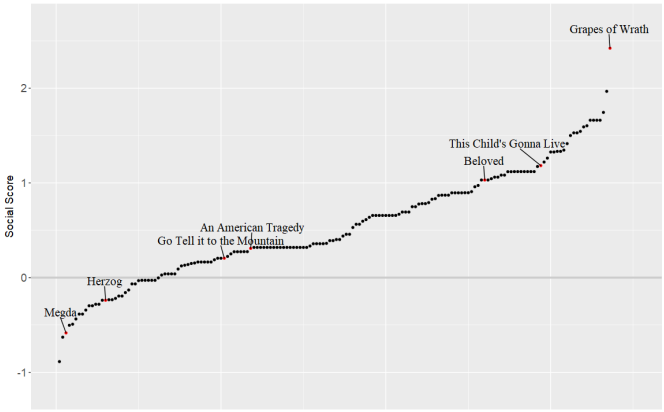

Figure 3. Plot showing the "social" score for all novels containing alignments with the Bible. Lower scores indicate novels where the Bible is less frequently cited in a "social" way, as we define the term. Scores closer to zero indicate novels where the "social" and "non-social" contexts are split evenly, as in James Baldwin's Go Tell it to the Mountain.

According to these scores, one of the more "social" novels in our corpus is Sarah Wright's *This Child's Gonna Live*, from 1969. Of the twelve biblical quotes in the work, eleven are made in a "social" context. They form part of a longer conversation happening in the novel between its protagonist, Mariah Upshur, and the church. A twenty-three year old black woman living a subsistence life in 1930s rural Maryland and pregnant with her fifth child, Mariah, as Trudier Harris writes, "must reevaluate and ultimately reject Christianity" to challenge the social and institutional forces that deny her individuality and autonomy.[38] She does so by imagining Jesus and God as characters in her mind with whom she can converse and argue, closing the gap between herself and the divine. She reduces "Jesus and

[38]Trudier Harris, "Three Black Women Writers and Humanism: A Folk Perspective," in *Black American Literature and Humanism*, ed. R. Baxter Miller (Lexington, KY: University Press of Lexington, 1981), 55.

God to the role of conversational buddies… Jesus becomes a familiar companion who is addressed without reverence and who perhaps takes the place of Mariah's nonexistent friends."[39] At first glance, Wright's novel validates the "critical modification" thesis and our statistical model. Yet it does so in a way that highlights the intersectionality of race and gender as crucial to particularizing that thesis.[40]

Mariah's socio-economic position hints at another intersection that is missing from our model. Across the entire corpus, John Steinbeck's *The Grapes of Wrath* is the most "social" novel. Like Wright's novel, it too centers on an impoverished rural family, and yet it also stands as an outlier in being by a white identified author. The novel gets ranked as highly "social" because of a pivotal scene in which Tom Joad, speaking with his mother, contrasts the "hell-fire Scripture" that tells the poor to endure their fate on earth with snippets of verse given to him by Jim Casy, a former preacher who has lost his faith. The snippets are reinterpreted by Tom as incitement to work toward a better life and a newly autonomous community in the here and now.[41] The scene embodies what Tamara Rombold identifies as a major structuring principle of the novel: "the rejection of the agency of God and… of traditional Christianity" through its "repeated inversions of Biblical material and the theological implications of those inversions."[42] To the extent we can align this structuring principle with the one that animates Wright's novel, we have here the seeds of an interpretive framework that reads "critical modification" of the Bible as motivated by intersecting forces of oppression along racial, but also gender and class-based lines. Viewed in relation to one another, the political project of narrating resistance to the authority of God's word is no longer a strictly "white" or "black" one.

This breakdown in categorical thinking occurs again when we look at a novel that is the inverse of Steinbeck's: one of the least "social" novels by a black identified writer. *Megda* (1891), by Emma Dunham Kelley, is a loosely autobiographical narrative that tells the story of a young woman, the titular character, who is studying to be a teacher in Rhode Island. An outlier in our model, it is also an outlier in literary history. As Holly Jackson has noted, *Megda* stands out from fin-de-siècle

[39] Harris, "Three Black Women Writers and Humanism," 55.

[40] The importance of intersectionality for reading Wright's novel is further attested to by what little critical attention has been given it. Jennifer Campbell celebrates the novel for its attempts to bridge "the movements for black Power and for women's rights." As she puts it, the novel "offers a woman-centered vision of an impoverished, besieged black community finally 'closing rank' in order to combat the systemic and individualized racism that seeks to destroy whatever community it cannot control." See " 'It's a Time in the Land': Gendering black Power and Sarah E. Wright's Place in the Tradition of black Women's Writing," *African American Review* 31, no. 2 (Summer, 1997), 212.

[41] John Steinbeck, *Grapes of Wrath* (New York: Penguin, 2014), 440-41.

[42] Tamara Rombold, "Biblical Inversion in 'The Grapes of Wrath,'" *College Literature* 14, no. 2 (Spring, 1987), 147.

black novels in its seeming unconcern for African American life or politics. It follows "a group of carefree adolescent female friends through Christian conversion to appropriate wifehood" and contains "no mention of the hardships facing black people at this time." Even more unusual is "the apparent whiteness of [Kelley's] characters, most of whom have blue eyes and skin described repeatedly in comparison to 'pure' or 'driven' snow." And yet, Jackson adds, "Kelley has been perennially cited and studied as an example of the diversity of strategies employed by women of color in the 1890s."[43] These strategies are read through the apparent fact of Kelley's identified race, making them part of, as Phillip Brian Harper puts it, "the peculiar effects of African American culture's having been conceived as a political project." This concept, however, also harbors within it the possibility that "any given work—not to mention the artist who produced it—is always liable to be deemed not properly black."[44]

If we look at Kelley's specific strategies for citing the Bible, we find she diverges sharply not only from contemporary peers, but from black writers across the twentieth century. Of the twelve Biblical citations found by our sequence alignment algorithm, we coded nine as "non-social." Many of these appear in the context of sermons delivered by Mr. Stanley, a preacher at the church attended by Megda and her family and friends. Some come in the middle of a pages-long monologue that lulls the reader into feeling as if they are but one of the congregation. Others come at the end of sermons and precede reactions by the congregation or by Megda herself. In the novel as a whole, her spiritual awakening provides a narrative focal point and performative template for how to submit to God's will. In one pivotal scene, in which Megda watches a baptismal service for her wealthier friend (and spiritual guide), Ethel, we see that the proper way to receive his message ranges from passive acceptance to rapturous attention. During the service, Mr. Stanley's "especially good" sermon leaves "not many dry eyes among the congregation" (217) while "perfect stillness" settles over the congregation after a "short but impressive" prayer (223). Megda's own affective response, heightened by feelings of attraction to Mr. Stanley, further frame this collective response. "A quiver passed over Meg's face as she listened; her lips trembled and her eyes filled with tears" (216). Moments later, "as she listened to the earnest words that came from his lips, she felt her heart throb and beat with a feeling she had never experienced before" (217). Even without Mr. Stanley's "beautiful face" to guide her, however, Megda was primed, with help from the narrator, to receive God's word. In the scene leading up to the baptism, Ethel admonishes

[43] Holly Jackson, "Identifying Emma Dunham Kelley: Rethinking Race and Authorship," *PMLA* 122, no. 3 (2007), 729.

[44] Phillip Brian Harper, *Abstractionist Aesthetics: Artistic Form and Social Critique in African American Culture* (New York: New York University Press, 2015), 1.

Megda for her reluctance to commit herself to the authority of God by reciting a few lines of scripture and leaving her to ponder them. Megda "stood there alone — speechless, remorseful, and dismayed, but, thank God! no longer blind to her own folly and wickedness" (203-204).

These scenes of Biblical quotation convey the antithesis of critical modification, lacking any sense of irony, critique, or dialogism. There appears to be a major category error in identifying Kelley as "black" in light of the associations between race and critical modification uncovered by qualitative accounts and reinforced by our computational model. Investigating the matter further, we found that until 2007, little was actually known about Kelley's biography. *Megda* was itself mostly forgotten until 1955, when the bookseller Maxwell Whiteman listed it in a landmark chronology, *A Century of Fiction by American Negroes 1853-1952: A Descriptive Bibliography*. As Katherine Flynn notes, Whiteman assumed, like many after him, Kelley-Hawkins's African heritage from a frontispiece photograph in the novel (Figure 4).

Figure 4. The portrait of Emma Dunham-Kelley included in the frontispiece to Megda (1891).

In 1976, Kelley's status as a "Negro author" was solidified with *Megda*'s inclusion in the catalog of the Schomburg Center for Research in Black Culture. A decade on, Gates included her in his Schomburg Collection of African American Women Writers of the Nineteenth Century series, which he was inspired to compile after discovering a second novel by Kelley. "For a time," Flynn writes, "Kelley-Hawkins was understood to be the first published African American female novelist, prompting several studies of her work—the majority of which examined and tried to rationalize the absence of African Americans and racial issues."[45] Critics argued that Kelley's original audience of black readers would have interpreted her female characters as "white mulattas," or construed the exclusively white world of

[45] See Katherine Flynn, "Emma Dunham Kelley-Hawkins 1863-1938," *Legacy* 24, no. 2 (2007), 283

her novels as "a kind of postracial utopia."[46] As Jennifer Harris writes, there has been a "willingness to accept the irreconcilabilities, elisions, and oddities" that come up when reading race in Kelley's writing because of "the very way that we read African American literature as always playing with… such matters." Such interpretive machinery has meant that Kelley is seen to be subverting "white literary codes" to covertly advance "black causes."[47]

A major wrench was thrown into this machinery in 2007, however, when Flynn and Jackson independently confirmed via historical records and census data that Kelley did not identify as a person of color at any time in her life.[48] Moreover, "every one of the official records designates Kelley and every member of her family as racially white."[49] In the wake of this discovery, it became clear that Kelley's assumed racial identity had been a cipher through which critics were reading the play of racial difference in her work. She was never passing as "white" because, according to these records, she had no need to. Yet this cipher inadvertently entered into our model owing to its social accretion through a series of editorial decisions by African-American literature scholars. Sticking to the grammar of comparison, one response to this situation would be to change Kelley from "black" to "white" in our metadata and declare our initial assignment as a failure to acknowledge this recent discovery. Indeed, this is effectively what Gates did when he learned of the discovery from Jackson, announcing that Kelley's novels would be withdrawn from the Schomburg series.[50] They were, in Harper's words, "not properly black."[51] Oddly, Gates himself, in the introduction to his series, wrote that "Literary works configure into a tradition not because of some mystical collective unconscious determined by biology of race or gender, but because writers read other writers and *ground* their representations of experience in models of language provided largely by other writers to whom they feel akin."[52] But in deciding to remove Kelley from the series, he implies that biology does matter: as a white woman, Kelley could not possibly have felt kinship with the black woman writers of her day. Biology begets social affinities begets patterns of literary relation. Gates excludes her based on categorical thinking.

Rather than double down on this coupling of racial identity to aesthetics, we

[46] Jackson, 729.

[47] Jennifer Harris, "Black Like?: The Strange Case of Emma Dunham Kelley-Hawkins," *African American Review* 40, no. 3 (Fall, 2006), 414.

[48] See both Jackson and Harris.

[49] Jackson, 730.

[50] Jackson, 739.

[51] Interestingly, Jackson herself absolved Kelley's novels of having any political thrust in light of their depiction of an extremely white world, but also presumably in light of what she had discovered about her racial biography. See Jarrett, 318

[52] Cited in Jackson, 738.

want to think about *Megda*'s outlier status through the grammar of relationality. This means thinking about it in relation to other texts through the "models of language" they share - in this case, models for situating personal identity (as a narrated process) within religious practices and institutions. While the language of critical modification appears absent in the novel, this is in part because our computational procedure accentuates those moments where direct citation of biblical verse leaves Megda and others speechless and awestruck. They are given no choice but to submit to the authority of God's word. Blind allegiance to the power of an Other, however, is predicated on the repression and denial of other sources of authority. *Megda* does not expel these sources outside the text itself, but rather embeds them within the titular character's own internal psychological struggle, one that animates the narrative leading up to the baptism scene. As the narrator frames it, Megda's struggle is one of a lost and confused soul who awaits someone to lead her out of "the tangled path of doubt and dark uncertainty in which she was walking and place her feet in the narrow, shining way" (67). The familiar Christian tropes of darkness and light, enslavement and freedom, operate throughout to remind us of where Megda walks and where she must go, even marking her physically with "sparkling dark eyes" (36) in contrast to Ethel's "transparent," "fair," and "very white" skin (108). There is never any doubt she will find her way. Narratorial interjection and the characterization of everyone *but* Megda aim the text's ideological crosshairs steadily on conversion. But if the novel on the whole opposes critical modification of Biblical authority, it still needs Megda to show what resistance (and inevitable submission) look like.

According to the narrator, the tyrant "pride" is Megda's "besetting sin," keeping her from realizing she needs "a Saviour's help" (102). But as narrated in her own mind, her resistance stems from not wanting to be told what church to join (25); not wanting to lower herself in her "own or anyone else's estimation, by making false professions of religion" (30); not wanting to "be governed entirely by one superior mind" or to be one of those people who "never give an original expression to an original idea" (61). She feels this resistance no less outside the church than within, and is first portrayed as someone who, unlike the rest of the congregation, cannot "keep her thoughts from dwelling, first upon one thing then upon another, instead of keeping them strictly upon the sermon" (110). Indeed, "the beautiful words from God's own book, the grand thoughts [Mr. Stanley] gave expression to… were all lost upon Meg; she gave no thought to them" (113). Perhaps her biggest hurdle to accepting God's will is her love for theater, a passion that plays out prior to her spiritual awakening in the form of a school recital, performed for the whole town and in which she takes center stage. It is hardly ironic that her solo performance in the recital—a reading of a poem about a mother willing to have her arms and hands bound by "Russia's heaviest iron bands" to save

her son's life—is met with a reaction identical to Mr. Stanley's sermons: "the audience showed their appreciation of her effort in that truest praise of all—complete, breathless silence" (155).

Ultimately, it is Megda's charisma that stands as her biggest threat to the novel's ideological orientation. And it is from this point on that the "workings of the Spirit" begin to make her "a little more thoughtful, a little more subdued in her manner" (177-78). When she hears Ethel's voice in prayer she is now transfixed (190); Mr. Stanley's prayers compel her to give her "whole mind" to them (192). At the pivotal moment following Ethel's baptism where Megda finally, inevitably, crosses into the light, the text reinforces how complete is her subjugation to the authority of God and the sacrifice of her own freedom to resist. When Mr. Stanley relays an anecdote about a desperate Siberian prisoner whose chains are miraculously unfastened by Jesus, representing the latter's sacrifice for the spiritual freedom of all (227-230), Megda has her long awaited conversion. She "felt like falling on her knees then and there, and bowing her head before the power of the Omnipotent" (230). The experience leaves her "inexpressibly happy," and so too, we can imagine, the intended reader, now that all possibility of critical modification is suppressed. Even the narrator steps in to remind us that, "all this peace and joy would have been hers months ago if she had only opened her heart to Him, surrendered her will to His and believed on Him as her Saviour" (231).

At this final surrender, with Megda relinquishing one set of chains (of pride) for another (to Jesus), the novel reconfirms its overall ideological stance. Moreover, it firmly situates its rejection of critical modification as a movement toward whiteness, silent submission, and, by the end of the novel, safe middle-class domesticity. *Megda*, situated at the extreme end of the general narratological patterns revealed by our computational model, is thus right where it should be—its attitude to the Bible is decidedly "non-social" and, moreover, conditioned on "whiteness" being an amplifier of this orientation. We can almost read the text as a whole as confirmation of the categorical thinking that allowed us to situate it within these broader patterns in the first place. "White" writers tend to engage the Bible in one way, "black" writers another. And yet we found that, according to this thinking, *Megda* was also *not* where it should be. We could have explained this away as an effect of the author's curious biography and the series of decisions that led her to be improperly slotted into the category of "black" identified author. Instead, we treated the novel as the very exception that proved the possibility of thinking past whether Emma Dunham Kelley was writing while "white" or "black." This meant reading it against the same general patterns found by our model, but this time through a relational perspective wherein social and non-social attitudes toward the Bible are seen as part of a continuum, neither divided along strict racial

lines nor fully determined by them.

We were, in other words, able to read the novel doubly. For when recognized as part of a broader formal tendency to appropriate the Bible in a non-critical way, we can read it alongside novels that exhibit this tendency irrespective of the racial identity of their authors. Conversely, we can read together novels showing the opposite tendency, just as we began to do with *This Child's Gonna Live* and *Grapes of Wrath*. These new kinds of groupings allow us to read the interaction of race, writing, and religion across a set of shared formal tendencies - shared "models of language" - and not across the categorical labels provisionally assigned to authors. In the case of *Megda*, our relational approach further revealed how the continuum between social and non-social attitudes - itself structured by specific interactions of race and religion with other dimensions like gender and class - could play out within the individual text or even within individual subjects. Race, understood as an assemblage of social forces and institutions that exceed the ontological and categorical division of human experience, can be read everywhere in the US novel, however particular its manifestation may be in any one text.

To conclude, and to return to our initial gambit of bringing together computation and the critique of race: where does this leave us? Our approach allowed us to critique an obvious and easy target: Northrop Frye and his theory of the Bible and literary universalism. Yet, it also enabled us to revisit an important controversy within African-American literary scholarship, and to expose the limits of categorical thinking as it appears in quantitative methods, but also in one strain of humanistic research. Counter intuitively, computation helps us think beyond a politics of identification that, as in Gates's editorial decision, can easily fall back to a flat mapping of racialized bodies onto racial texts. The key is focusing on both what the method reveals at scale as pattern, but also transforming our perspective on that pattern by questioning its underlying assumptions. For us, that meant substituting the grammar of categories for the grammar of relationality to read the general patterns doubly, and so too the individual texts, like *Megda*, that constitute them. We still need categorical labels to limn the broad contours against which to orient the interaction of race and religion across the twentieth-century US novel. Abstract and reductive as they are, they point to a fundamental dimension around which social and literary life is organized. But if this dimension provides an initial means of seeing difference in the data, it is one we can look beyond. It was our double vision of *Megda* that allowed us to loosen the categorical thinking behind our method, but also behind literary history itself, which in this particular case, uncannily echoes the former.