HTRC/ACS Final Report
Dan Sinykin

Overview
Sixty years ago, most publishers in the US were independent. In subsequent decades, multinational media conglomerates consolidated the field, such that by 2000, 80% of fiction was published by just six conglomerates. How has conglomeration changed literature?

People argue either that the changes have been bad for literature or they have been great. Either conglomeration has ruined literature by submitting it to the demands of the bottom line, constraining the ability to publish good literary books, or conglomeration has suffused the industry with capital, allowing for a flourishing of diverse literature.

In *The Conglomerate Era*, under contract with Columbia University Press, I take this impasse as the occasion for an inquiry into conglomeration's effects. Synthesizing archival research and computational analysis, I move beyond demonization and apologetics to develop a theory of American literature in the conglomerate era—a period that lasted from Random House's decision to go public in 1960 to the financial crisis and the release of the Amazon Kindle in 2007.

Project
How did the conglomeration of the publishing industry change literature? Answering this question would be impossible without computational analysis—the scale of contemporary literature is too large, otherwise—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.

I opened an account with the HTRC in September 2018, upon learning that it made data capsules and virtual machines available for public use. I developed publisher-based corpora in my capsule: one of Random House novels, another of small nonprofit press novels. I used predictive modeling (text classification via logistic regression) to determine whether Random House and nonprofit novels can be distinguished, and by which features. Nonprofits claim that what distinguishes them from commercial presses is literariness, so this modeling tests that claim.

But my research was limited by computing power and my modest abilities. The ACS program provided the assistance necessary for me to take full advantage of the HTRC's rich offerings. With the support of the HTRC, I

- **built publisher-based full-text corpora for ten presses;**
- **helped develop a program to clean text of headers and footers;**
- **replicated my study of differences between Random House and nonprofit novels;**
- **and designed and completed an experiment to study the differences between trade and mass-market novels**.

I will devote the remainder of this report to my findings on the last study: whether there are measurable differences between trade and mass-market novels, identifiable by a machine-learning classifier.

Methods

I relied on HTRC staff support to build publisher-based corpora. Ryan Dubnicek performed SQL queries on the HathiTrust holdings to build corpora for eight publishers, which formed the basis of my subsequent work. I relied, too, on HTRC staff support to develop a program to clean these texts of headers and footers—a program that will be invaluable to future researchers. I paid a student to hand clean the texts of front and back matter.

The machine-learning model I then used to differentiate between trade and mass-market novels employs regularized logistic regression with an L1 penalty. I calculated accuracy from the results of leave-one-out cross validation. I used Python to code the model, and the scikit-learn library to conduct logistic regression. Features are words that have been transformed into doc-term matrices for calculation. I took several steps to determine which features are salient in distinguishing classes, normalizing the model, performing a Ztest, and determining p-values. I included features that pass a p-value threshold of 0.05.

Ted Underwood provides an extensive account of logistic regression for predictive modeling of texts in Appendix B of *Distant Horizons*. I also hew closely to the practice of Richard Jean So and Edwin Roland in "Race and Distant Reading."

The full code is available at: https://github.com/sinykin/mass-market.

Outcome

What makes mass-market fiction different from trade? Format: mass-market books are smaller and printed on cheaper paper. Distribution: mass-market books are produced in much higher numbers and disseminated through airports and superstores (though this long ago became true of some trade, too). How about what's in them? The content?

Mass-market books are frequently trade reprints. And mass-market originals are often printed as trade, too. Many mass-market publishers created trade lines for that purpose. So are they the same books just packaged and sold differently? That's not how we tend to think about it. We tend to see mass-market books as beach reads, schlocky entertainment, vehicles for escape. In practice, mass-market books are those that publishers think they can sell to vast audiences. Trade, in contrast, can exist for a niche.

I built collections, or corpora, of fiction from four major mass-market publishers (Bantam; Dell; NAL; and Pocket) and four major trade publishers (Farrar, Straus, and Giroux; Harper; Random House; and Simon & Schuster). These corpora span 1950 to 2007. I taught a computational model to learn how to distinguish between corpora. We give the model a batch from each category to train on; as it processes the two categories of text, it discovers features—in this case, words—that tend to help it predict to which category any given text belongs. After we've trained the model, we give it texts without labels and ask it to guess its category. We learn, then, how well-defined the categories are. If they're indistinguishable, the model will guess correctly 50% of the time. If they're easy to tell apart, the model will guess correctly with the proficiency of a spam filter, better than 90%. The model records the weights it gives to individual words, so we learn what exactly characterizes the categories in relation to one another. Code and results are available at github.com/sinykin.

I did several tests. How different are mass market and trade books from one another across the full span of years? How different are they between 1950 and 1980, and do they become easier or more difficult to distinguish between 1980 and 2007? And if I trained a model to distinguish based on the earlier period, how well could it perform on corpora from the later period?

The model ought not be good at telling the difference, given the considerable overlap between the two categories. It's not. Across ten runs on random subsets from the full chronology of the corpora, its precision ranged from 59.8% to 68.5%. This is about what I expected. Insofar as the model can distinguish between them, it notices that mass-marked fiction is more likely to use adverbs and past tense: breathlessly, doubtfully, dryly, hastily, mercifully, ominously, smugly, speculatively; chuckled, demanded, glanced, grinned, shivered, shrugged, snorted, tightened. These words also confirm my expectations. We recognize that adverbs like ominously and smugly are more likely found in the popular fiction favored by the mass market, as are these specific past tense verbs, which suggest melodrama. The model finds that present participle and present tense help identify trade fiction: fussing, gusting, smoking, swooning, wheezing; humiliate, listen. The few past tense verbs that help identify trade differ from those for the mass market: announced, enjoyed, introduced. We begin to get a sense that the model sees between the categories a class difference: trade fiction eschews the blatant expressiveness of adverbs for restraint; characters in mass-market fiction demanded ominously whereas those in trade listen, fussing and swooning.

The model distinguishes the categories with 68.3% to 72.0% accuracy between 1950 and 1980; that range declines to 60.5% to 66.1% for 1980-2007. This might at first seem surprising. Didn't mass-market publishers publish fewer prestigious books? Less Faulkner, more Spillane? Less Morrison, more Krantz? Not really. Mass-market imprints still published prestige so long as it sold: Atwood, Morrison, Munro, and Pym all show up in the corpora with mass-market books. And brand authors show up on both sides, too: Crichton, Higgins Clark, and Turow appear just as trade; King and Koontz as both; and Auel, Conroy, Follett, and L'Amour just as mass-market. No wonder the model gets confused.

Institutionally, conglomeration brought mass-market and trade publishers together under a single roof. Both were asked to submit to the bottom line. At both, the autonomy of editors declined. The results from the model suggest a homogenization, with both moving together in the same direction.

If mass-market and trade fiction move in the same direction after 1980, what direction is that? To find out, I used a method developed by Ted Underwood, which he calls perspectival modeling. I trained a model to distinguish between mass-market and trade fiction between 1950 and 1980 and tested its ability to differentiate between the categories between 1980 and 2007. The model adopts the perspective of the earlier period and evaluates on that basis. The results were startling. It appeared that the model was somewhat less precise than that which studied the full span from 1950 to 2007, ranging from 53.8% to 61.5%. But when I investigated how well it did with each category, I was surprised. It was exceptionally imprecise at guessing mass-market fiction, only getting it right 30.3% to 41.5% of the time. Meanwhile, it was quite good with trade fiction, correct 66.7% to 83.6% of the time—with 66.7% as a low outlier, the next lowest being 76.7%. With the past as its guide, the model improved dramatically in its ability to recognize trade fiction. And it mistook mass-market fiction for trade fiction considerably more often than not.

The results confirm that mass-market and trade moved in the same direction: away from the former and toward the latter. On what basis? Past tense gave way to present participle and present tense. And nouns became more central. Much is as above. The obsolescent past tense verbs tended toward brusqueness and physicality: blasted, burned, clung, jerked, prowled, shook, slammed, thrust. The incipient verbs suggested social mores and subtler expressions of power: apologize, arranged, discourage, divorced, folding, imposed, recognizes, remembers. The rising nouns corroborate this view: alabaster, austerity, Connecticut, gestures, goodbye, impressions, salary, sermon, slum, Yorker. From a world where people prowled and thrust, we've entered one where they apologize for their salary and discourage squalid impressions in Connecticut.

I've arranged the terms hyperbolically. But the effect is real and counter-intuitive, at least for those of us who believe that fiction has become increasingly commercial under conglomeration. Were that the case, we might conclude that fiction would move in the direction of the mass market. Not so. From the perspective of a longer timeframe, the trend toward trade is surprising for a different reason. As noted above, Ted Underwood has shown that across centuries, English-language fiction has moved away from abstraction and politics and toward language of embodiment and perception. After 1980, fictional language became less bluntly embodied. While still concrete and perceptual, it became more mental than physical. It moved it the direction of subtler experiences.

We need to take these results with a grain of salt. For one, we are not seeing differences between the full lists of mass-market and trade publishers, but subsets of those lists that were deemed worth holding by the university libraries from which HathiTrust built its collection. Despite the fact that these holdings include major brand names such as Auel, Crichton, King, and Koontz, we might expect libraries to privilege prestige over popularity, skewing the results. It's also possible that the process of digitizing (scanning, OCRing) mass-market books works differently from trade because of font size and paper quality. Such investigations are beyond my purview. It is enough, here, to take these provisional results as additional evidence for understanding the trajectory of the mass market in the latter half of the twentieth century.

Appendix: Links

For further details on the results of my work with the HTRC, visit the following:

- https://www.publicbooks.org/how-capitalism-changed-american-literature/