Book Reviews and the Consolidation of Genre

Kent Chang, Yuerong Hu, Wenyi Shang, Aniruddha Sharma, Shubhangi Singhal, Ted Underwood, Jessica Witte, Peizhen Wu

A paper presented at the virtual panel "Cultural Analytics and the Book Review: Models, Methods and Corpora," *ADHO 2020*, July 22, 2020.

Introduction

Book reviews clearly cast new light on reception: on literary judgment, for instance, and prestige. But reviews may also give us an opportunity to test claims about the significance of patterns in the reviewed books themselves.

For instance, literary scholars have recently claimed that predictive models can measure the strength of the boundaries that separate different cultural categories—different genres of fiction, say, or market segments.[1] But the evidence supporting this argument comes purely from the texts themselves. The works in a particular literary genre may be relatively easy (or hard) to distinguish from others, because they possess (or lack) a distinctive diction. Interpreting this textual boundary as evidence about the strength of a cultural distinction has seemed questionable to many readers.[2] One can imagine cultural categories that would be salient and distinctive for human readers even though they don't leave the kind of traces that can be captured in a model of word frequency.

So do textual models really tell us anything about the boundaries between cultural categories? It is hard to resolve this question with a single experiment, because it isn't immediately clear what counts as ground truth about the strength of a cultural boundary. José Calvo Tello has compared predictive accuracy to the *level of human consensus* about different genres (expressed, for instance in bibliographies).[3] Ted Underwood has

[1] Dan Sinykin, "How Capitalism Changed American Literature," *Public Books,* July 17, 2019, https://www.publicbooks.org/how-capitalism-changed-american-literature/. Richard Jean So and Edwin Roland, "Race and Distant Reading," *PMLA* 135.1 (Jan 2020): 59-73.

[2] This is, for instance, one of the questions raised by Nan Z. Da, "The Computational Case against Computational Literary Studies," *Critical Inquiry* 45 (Spring 2019): 601-39.

[3] José Calvo Tello, "Genre Classification in Spanish Novels: A Hard Task for Humans and Machines?" *European Association for Digital Humanities* 2018, https://eadh2018.exordo.com/programme/presentation/82.

compared predictive accuracy to *the degree of overlap or separation* between genres. (That is, we might expect pairs of genre labels that are often assigned to the same works to be closer to each other than those that rarely overlap.)[4] Both studies suggest that the accuracy of a textual model does correlate with the behavior of human observers. But both studies are still open to the objection that they rely purely on explicit labels. This could produce a subtle kind of false confirmation. Perhaps the *conscious* labeling behavior of bibliographers and catalogers is governed by categories overtly signaled in the diction of a literary work—but ordinary readers care more, in practice, about other categories, less clearly registered in diction?

Book reviews give us a way to address this remaining source of doubt. Reviewers may or may not explicitly assign books to a genre: in the nineteenth- and early-twentieth-century period we will discuss here, explicit genre categorization is unusual. But reviews do presumably reflect the tacit concepts and categories that organize the landscape of fiction for a particular reader. It seems likely that books with similar reviews were perceived in similar ways. So if the textual boundaries between groups of literary works do really correlate with the responses of ordinary readers, reviews of those texts ought to reveal the same groupings and distinctions. That is the primary hypothesis we set out to test in this paper. Are the subject or genre categories most strongly marked in fiction also the categories most strongly marked in reviews of fiction?

Data

To construct a corpus of paired literary texts and book reviews we aligned extracted features from HathiTrust Research Center with book reviews from ProQuest's *British Periodicals* collection, matching on both the author and the title of the original work.[5] We also used predictive modeling to filter the book reviews for reviews of fiction. Review metadata is imperfect, and title matches are often ambiguous, so without this filtering step it would have been difficult to have confidence that we were really pairing books of fiction with their reviews.

[4] Ted Underwood, "The Historical Significance of Textual Distances," Proceedings of the Second Joint Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Santa Fe, 2018, https://www.aclweb.org/anthology/W18-4507/.

[5] Boris Capitanu, Ted Underwood, Peter Organisciak, Timothy Cole, Maria Janina Sarol, J. Stephen Downie (2016). The HathiTrust Research Center Extracted Feature Dataset (1.0) [Dataset]. HathiTrust Research Center, http://dx.doi.org/10.13012/J8X63JT3.

When the filtering process was complete we had 9137 pairs of books and reviews. The books' dates of first publication extend from 1535 to 1950, but the vast majority (more than 9000) were published after 1800, and about 8000 after 1850. The reviews dated from 1800 to 1950. Most were published very shortly after the book in question, although we do have a few nineteenth-century reviews of *Don Quixote.* When a book had multiple volumes, we aggregated the texts; when we had multiple reviews of the same book, we also aggregated the reviews to produce a single composite review-text. Word counts for the books and reviews are available through a GitHub repository documenting the project.[6]

Methods

Our overall hypothesis was (generally) that similar books will have similar reviews and (more specifically) that categories of books with closely-knit textual similarity will also have reviews that resemble each other closely. We preregistered an initial plan to test this hypothesis in two ways: using supervised predictive models (which have worked well for this problem in the past, but require relatively large groups of works), and using Word Mover's Distance.[7] Distance metrics are easy to apply to small groups of texts, and we hoped a distance-measuring approach would allow us to explore this question across a wider range of genres, including genres with few examples. While other distance metrics are more familiar, we thought Word Mover's Distance might be preferable for short texts like reviews, since it uses word embeddings rather than one-hot encoding and thereby produces a less sparse feature space.

We did find that our preregistered hypotheses were confirmed using Word Mover's Distance. For instance, to take the simplest example, we measured WMD between 500 random pairs of books and the corresponding pairs of reviews. We found a statistically significant relationship between the two measures ($r = .13$, $p < .01$). But regular cosine distance on the frequencies of the most frequent 2500 words showed an even stronger relationship in the same sample (r = .22, p < .00001). In subsequent

[6] Metadata for the books used in this experiment is available at our GitHub repository, https://github.com/tedunderwood/reviews/tree/master/bpo/corexperiment. Metadata for the reviews (and word counts for both the books and reviews) is available at the supporting *Open Science Framework* site: https://osf.io/a3749/.

[7] Yuerong Hu, Wenyi Shang, and William E. Underwood, “Book Reviews and the Consolidation of Genre: First Registration,” *Open Science Framework,* October 9, 2019, https://osf.io/j2ycz.
Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, “From Word Embeddings to Document Distances,” *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, 2015.

experiments, we also found that measuring category strength with cosine distance produced results that echoed predictive models more closely than WMD did. So we reverted to cosine distance in subsequent experiments. Since this is already the dominant distance measure in text analysis, we didn't feel there was a great risk of tailoring our methods to a particular sample or problem.[8]

As part of our preregistered hypothesis, we used metadata in contemporary libraries to define twenty-four genre or subject categories. These categories could be viewed as a source of anachronism (since they were mostly defined by librarians half a century or more after the publication of the original works). But the anachronism in question is helpfully orthogonal to the question explored here. In other words, we're not positing that these twenty-four categories are the *best* categories for English fiction 1800-1950, or that they precisely align with real divisions in literary culture. Instead we are asking whether the relative clarity of different categories (in the literary texts themselves) correlates with the relative clarity of the same categories (in reviews of the texts). To fully test this hypothesis, it might actually be good if some of our categories were anachronistic, and did fail to align clearly with real boundaries between literary practices.

We then tested our central hypothesis about genre in several different ways. First, we trained classifiers to distinguish literary works (and their reviews) from other works and reviews in our corpus. (We used the scikit-learn implementation of regularized logistic regression.)[9] We found that the accuracy of the book classifiers correlated with the accuracy of the review classifiers, $r = .867$ and $p < .001$. In this part of the experiment, we could only use fourteen large categories, because predictive models become unstable with small training sets. So some of the categories in figure 1 are all-encompassing (e.g. "random"), or very general (works labeled "novel" or "romance" in their titles), or defined through subject headings rather than genres (e.g. works about "Britain" or "North America").

[8] Instead of using tf-idf, we scale features by converting each column of the matrix to a z-score, which is equivalent to using Burrows's Delta. For empirical evidence that this distance measure works well for many problems in text analysis, see Stefan Evert et al., "Understanding and Explaining Delta Measures for Authorship Attribution," *Digital Scholarship in the Humanities* 32.2 (December 2017): pp. 4-16, https://doi.org/10.1093/llc/fqx023.

[9] "Scikit-learn: Machine Learning in Python," Pedregosa et al., *JMLR* 12, pp. 2825-2830, 2011.

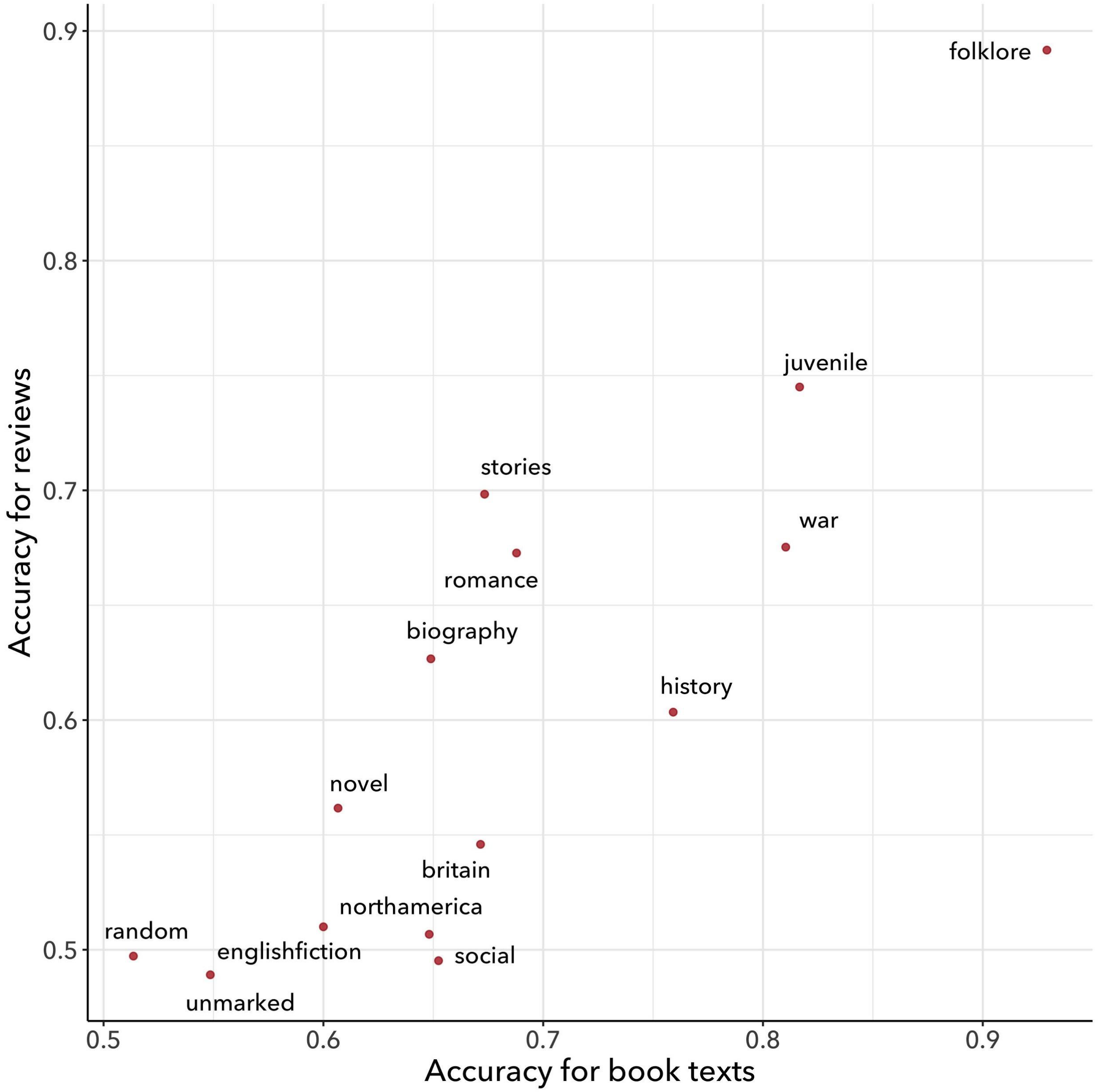

Figure 1. Correlation between predictive models of books and of reviews.

But we also observed the same pattern across a larger set of categories that are closer to groups ordinarily called "genres," using distance measurements between pairs of texts. We selected random pairs of works in the same genre and measured both the in-genre distance (between Mystery A and Mystery B) and the out-of-genre distances (e.g. from Mystery A to a randomly selected work published in the same year as Mystery B). Distances were measured as cosine distances for the 2,500 most common words, scaled using Burrows's Delta (which is in effect the StandardScaler in scikit-learn). By subtracting the in-genre distance from the out-of-genre distance for each pair, we obtained a measurement of how much closer works in each genre are to each other than to the rest of the corpus. Again, we found that closely-knit genres produce closely-knit groups of reviews, $r = .806$, $p < .001$.

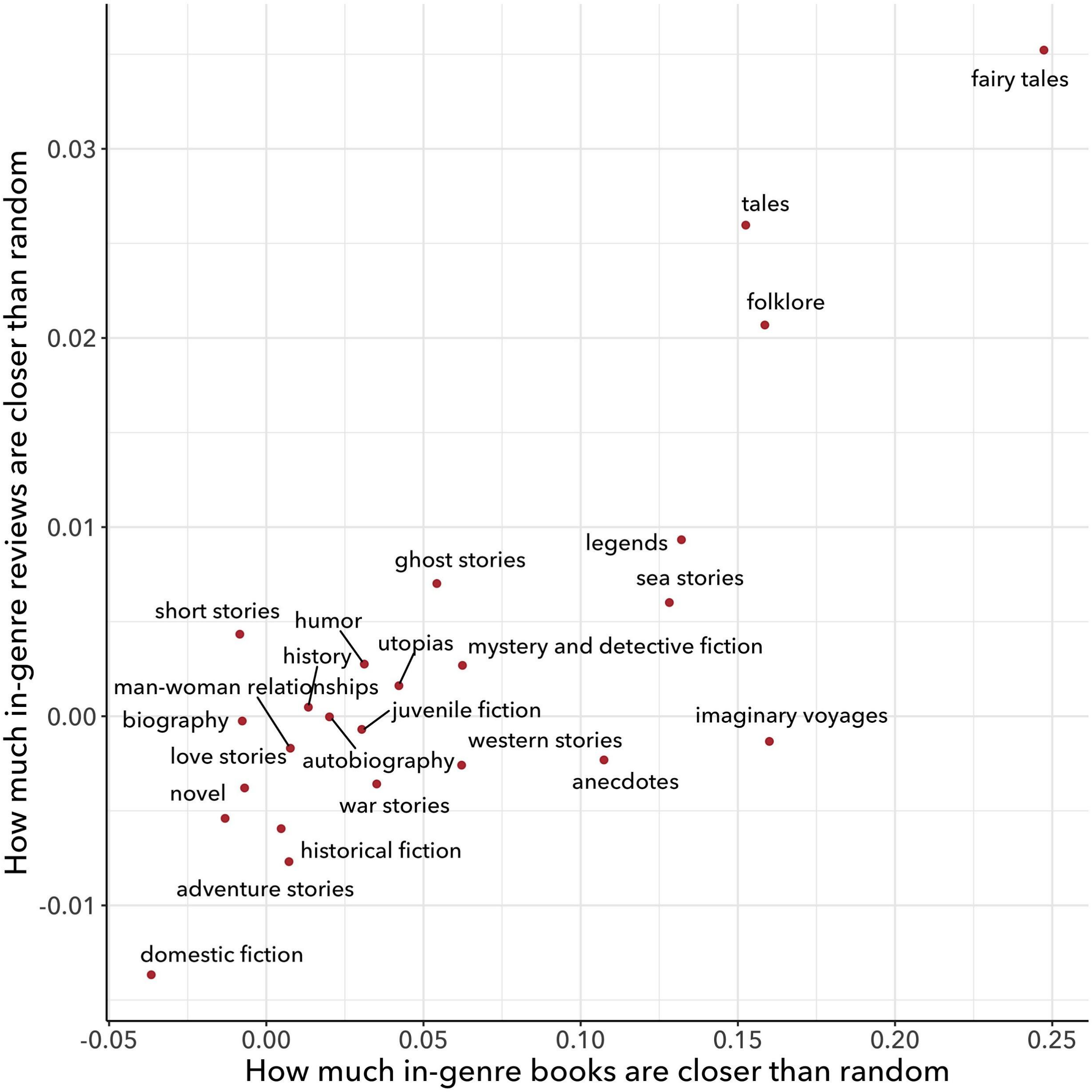

Figure 2. Correlation between distance-differences for books and reviews. In each case a pair of books in the same category are cross-compared to a pair of books outside the category, but published in the same years.

Conclusions and future work

We conclude that the similarities and differences between texts (measured, for instance, by cosine similarity) do correlate with similarities and differences in reception—or, at any rate, in book reviews. When we look at individual pairs of books, the relationship may not be very strong; perhaps $r \approx .22$. But if we back up, and gather books in categories, the aggregate relationship is stronger. Closely-knit genres also produce clusters of closely-related reviews.

This could be to some extent a verbal accident, if book reviews were distinguished by exactly the same unusual words overrepresented in the book texts. For instance, we might imagine that references to "valor" and "surrender" would characterize war stories (as well as reviews written about them). But inspection of the most strongly marked categories in our corpus does not lead us to credit this explanation. For instance, books are likely to be folklore when they mention *fairy, witches,* or *invisible*; those are some of the strongest features in our predictive model. But reviews are likely to be *about* folklore when they mention *traditions, collected,* and *popular.* In both contexts, folklore is marked by a distinctive diction—but it is a different diction in each context. So we suspect the coherence of these categories is not a purely verbal accident, but reflects an underlying social distinction. Books of folklore are genuinely unlike other kinds of fiction; that social distinctiveness is reflected (in different ways) both in their texts and in their reviews.

We also tested this hypothesis in several other ways. For instance, we found that the correlation between review-similarity and book-similarity holds (more weakly, $r = .418$) even if we use two *different* subsets of works in each genre: one to test the similarity of book-texts, and a different, disjoint set to test the similarity of review-texts.

Having validated this measurement of generic distinctiveness, we then used it, experimentally, to measure a broad structural change in fiction between 1860 and 1920. We measured the difference between in-genre and out-of-genre comparisons, and dated each pair of books to the midpoint of the two publication dates (since we precisely matched out-of-genre comparisons to the dates of the in-genre pair, the midpoint date was always the same). We found that genres became more closely knit across this period: that is, works of fiction became more similar to other works in the same genre than they were to randomly selected works from the same publication year.

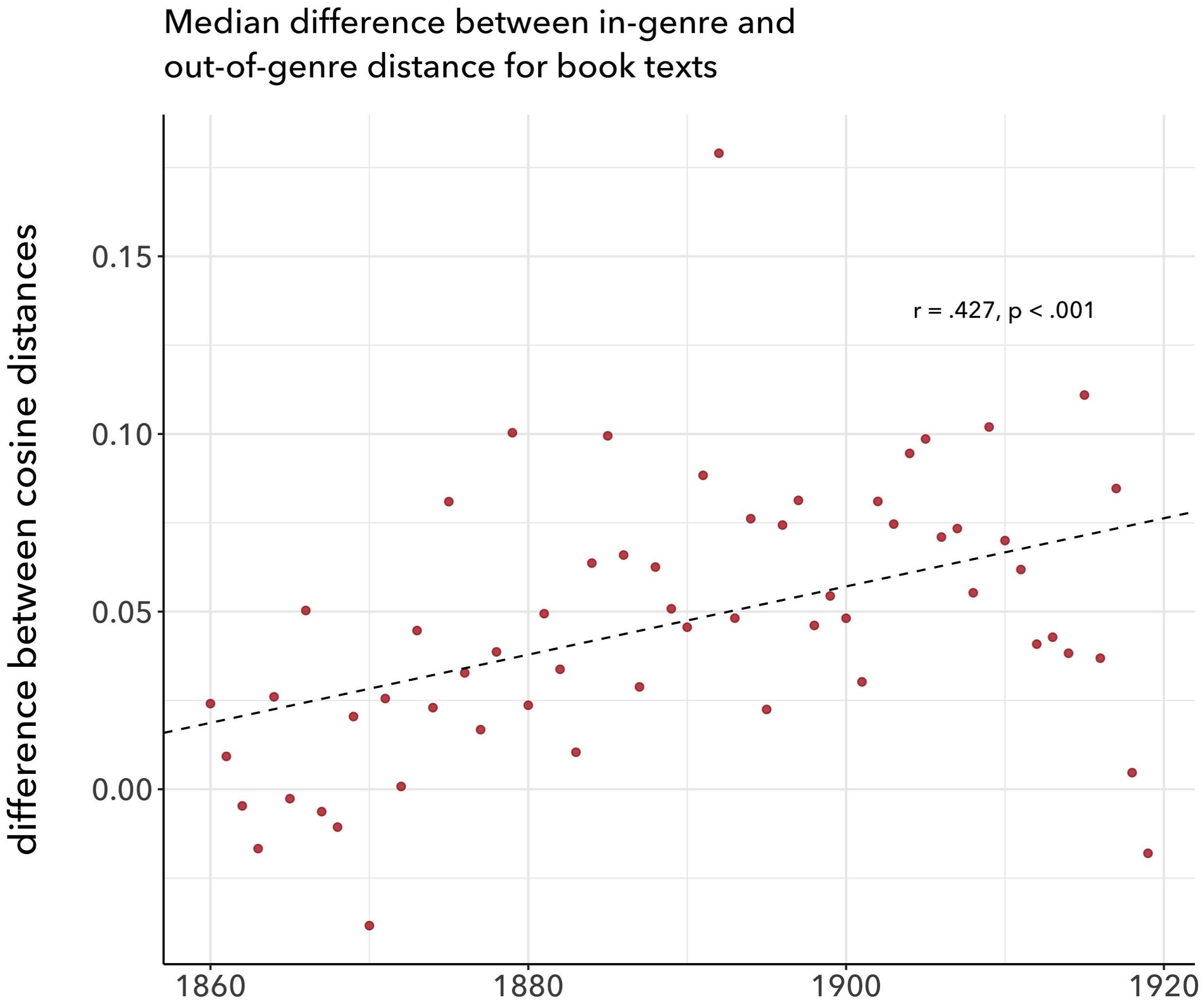

Figure 3. The consolidation of genre?

This pattern is open to different interpretations. It could reveal a process of differentiation (if we focus on the growing differences between genres)—or consolidation (if we focus on the strengthening of in-genre similarity). But since the categories we are using are drawn from late twentieth-century librarians' judgments, it could also be that works of fiction simply fit those judgments better as we move closer to the late twentieth century. To decide between these interpretations, we have subsequently repeated the experiment with categories inferred from a topic model of twentieth-century book reviews. But that's a different project and a separate paper.