The Rise and Fall of Genre Differentiation in English-Language Fiction

Aniruddha Sharma[a], Yuerong Hu[a], Peizhen Wu[a], Wenyi Shang[a], Shubhangi Singhal[a] and Ted Underwood[a]

[a] *University of Illinois, Urbana-Champaign*

Abstract

The organization of fiction into genres is a relatively recent innovation. We cast new light on the history of this practice by studying the strength of the textual differentiation between genres, and between genre fiction collectively and the rest of the fiction market, in a collection of English-language books stretching from 1860 to 2009. To measure differentiation, we adapt distance measures that have been used to evaluate the strength of clustering. We use genre labels from two different sources: the Library of Congress headings assigned by librarians, and the genre categories implicit in book reviews published by *Kirkus Reviews*, covering books from 1928 to 2009. Both sources support an account that has genre differentiation rising to (roughly) the middle of the twentieth century, and declining by the end of the century.

Keywords

genre, fiction, cultural analytics, English literature

1. Introduction

It has not always been intuitively obvious that fiction should be categorized by genre. In the nineteenth century, title pages do sometimes announce that a work is “a Romance” or “a Tale,” but those categories were apparently not important enough to organize books, for instance, in circulating library catalogs. Instead, nineteenth-century catalogs list books alphabetically by author [1, p. 45].

Around 1900, circulating libraries started to organize fiction under categories more responsive to content. But most of the new categories are better understood as subjects than as genres. That is, they describe what the text is about—not what it “is or what it does” [1, p. 63]. Organizing books by their geographical setting is especially popular, but subject categories can also be themes or plot elements (e.g., “jewel robberies”). Similar categories dominated the organization of fiction in the early-twentieth century Library of Congress [12].

The concept of genre was already familiar, of course: distinctions between epic, lyric, and dramatic poetry can be traced back to ancient Greece. But early-twentieth-century readers didn’t treat the topical differences between works of fiction as expressing an equally profound difference of kind. Subject categories hardened into institutions that are clearly “genres” only after pulp magazines started to divide readers into distinct markets with a confirmed preference for *Weird Tales*, *Ranch Romances*, or *Baffling Detective Mysteries*. The line between a subject

CHR 2020: Workshop on Computational Humanities Research, November 18–20, 2020, Amsterdam, The Netherlands

ashar29@illinois.edu (A. Sharma); yuerong2@illinois.edu (Y. Hu); peizhen4@illinois.edu (P. Wu); wenyis3@illinois.edu (W. Shang); ss100@illinois.edu (S. Singhal); tunder@illinois.edu (T. Underwood)
0000-0001-8960-1846 (T. Underwood)

CEUR Workshop Proceedings (CEUR-WS.org)

category and a genre can be blurry, however, and this process has been interpreted in divergent ways. Many literary historians would not say that subject categories turned into genres, but that the latent generic character of existing practices became more explicit. Historians often treat nineteenth-century stories about detectives, for instance, as a latent genre already equivalent to twentieth-century "Mystery"—even if readers around the year 1900 sometimes still divided books by asking whether they were about "jewel robberies" or "murders" or "Utah" [13].

In short, the emergence of genre in twentieth-century fiction is a long, muddled story, and it isn't clear that the conscious opinions of readers or writers provide decisive evidence about it. The lag time between literary practice and conscious categorization can be long. Different genres of fiction didn't commonly get labeled as genres in academic libraries until the last two decades of the twentieth century.

By that time, ironically, some of the genre boundaries that were finally getting formal recognition in libraries may have started to dissolve in literary practice. In 2011, for instance, Gary K. Wolfe observed that "Fantasy is evaporating . . . growing more diffuse, leaching out into the air around it, imparting a strange smell to the literary atmosphere" [22, p. viii]. Frederic Jameson has suggested that boundaries between genre fiction and mainstream literature may have begun to soften even earlier, since late-twentieth-century postmodernism entailed "the effacement … of the … frontier between high culture and so called mass or commercial culture," including "airport paperback categories of the gothic and the romance, the popular biography, the murder mystery, and the science fiction or fantasy novel" [9, pp. 2–3].

But have those generic boundaries really softened? If so, when? And did the postmodern transformation of genre allow mass-market genres to "diffuse" into mainstream literature, as Wolfe suggests? Or did it merely create a new, ironic high culture that incorporated fragments of downmarket genres like bottle caps in a mosaic? There is no agreement on any of these questions. It isn't even clear that we have the kind of evidence needed to address them yet.

We propose to cast new light on the history of genre by measuring the textual coherence of genres and genre-like subject categories in a large collection of fiction stretching from the late nineteenth century to the early twenty-first. How closely do works of fiction assigned to the same category actually resemble each other, and how does the strength of that clustering principle vary across the timeline?

Textual similarity is by no means the only kind of evidence needed for a history of genre. Many other things matter: how bookstores were organized, for instance, and how widely readers spread their purchases. But the strength of genre clustering in literary texts themselves is an important piece of evidence, and one largely missing at present. Moreover, we have reason to believe that this sort of textual evidence can illuminate other social questions about genre. Studies comparing textual models to the behavior of human observers have found that genres with strong textual boundaries tend also to be the subject of strong consensus among bibliographers and reviewers [4, 6]. So while textual evidence isn't the full story about genre, it should tell us whether genre differentiation has generally grown sharper or blurrier over the last 150 years.

2. Data

2.1. The general problem of historical representativeness

Our source for fictional texts is HathiTrust Digital Library [5]. Within this library, we relied on a subset of works identified as fiction in "NovelTM Datasets for English-Language Fiction"—specifically the "title list," which contains 138,164 volumes [19]. The outer boundaries of the NovelTM Datasets thus limit all the experiments that follow. For full discussion of those limits, see the NovelTM data publication. The two exclusions most relevant for this experiment are that HathiTrust is drawn mostly from academic libraries (so popular fiction may be underrepresented), and that the fiction datasets list book-length works rather than stories in magazines. Both gaps will tend to occlude pulp magazines—an important driver of generic specialization in the early twentieth century. It is thus possible that early-twentieth-century fiction was more strongly genre-differentiated than the NovelTM dataset, taken as a whole, would imply.

On the other hand, we didn't take the NovelTM dataset as a whole. The experiments we ran were based on subsets of books tagged as belonging to a genre by one of two groups of observers. This process of selection may well have tilted our data back in the direction of popular genre fiction. The net effect of these different selective processes, however, is also to encourage caution. The selection biases involved in historical data construction generally make it risky to interpret any subset of fiction as a random sample. Historians don't really have such a thing. So it is usually unwise to claim to determine any parameter absolutely for literature as a whole. On the other hand, it is still often possible to describe major trends. Empirical studies on the NovelTM fiction dataset have shown that the magnitude of historical change (across a long timeline) often outweighs the magnitude of the differences between different social groups or market segments [19]. However, there is no way to guarantee that this will be true for all questions. The best we can do is to construct multiple samples of the data, with different biases, and test a hypothesis in all of them until we find out where it breaks.

2.2. Sources for genre labels

Corpus linguists use the word "genre" for a specific dimension of textual differentiation. In one influential work by Biber and Conrad, the explicit formatting that defines a genre (for instance, the salutation in a letter) is distinguished from the pervasive linguistic coloring associated with a "register" or "style" [3]. This is a valid concept, but it's not the same thing literary critics or librarians tend to mean by "genre." Instead of inferring text types from the bottom up by observing their inherent characteristics, these other professions tend to treat genres as more-or-less conventional social categories. In recent decades, literary scholars have stressed that genres are "empirical, not logical" categories, which come into being "at particular historical moments," and may be defined differently by different observers [7].

Guided by this tradition, we approached genre as a perspectival attribute of texts. We ran two different experiments based on two different sets of observers.

- **(a)** In a first experiment, genres were defined by librarians (mostly located toward the end of the twentieth century) who assigned Library of Congress genre/form (or genre-like subject) designations. These are included with the NovelTM data.

  Of 138,164 volumes in the NovelTM title list, 19,365 were identified as bearing tags for one or more of the categories in Table 1. We randomly selected 7,081 volumes for our experiment—mostly from the volumes tagged with a genre category, although we also selected 1,125 volumes for a completely random contrast set (including volumes tagged with no genre at all).

When selecting the genre-tagged volumes, we mostly selected them in proportion to the genre's actual size in a given decade. (In other words, we allowed historical novels to be much more common in the nineteenth century than science fiction.) The exception to this rule is that we over-sampled smaller genres in order to ensure there were at least three volumes per genre per decade, whenever possible. 45 volumes were selected for this reason; over-sampling had a relatively small effect on the overall corpus.

- **(b)** In a second experiment, genres were defined by reviewers who described 19,018 books for *Kirkus Reviews* between 1928 and 2009. After topic-modeling those reviews, we selected certain topics as "genre-like." If one those topics was the dominant topic in a review, we assigned the corresponding book to a "genre." This produced a subset of 7,133 books assigned to genres.

Each of the sets of observers described above can produce a different kind of distortion. One problem with (a) might lie in the specific categories we selected as genres; see the discussion in section 3.3, "Genre-like categories." Another problem with (a) is that Library of Congress genre headings were not very commonly assigned before the last two decades of the twentieth century. So the tags in library metadata, mostly assigned rather recently, might represent a late-twentieth-century perspective on genre. If we found that the organization of texts into genres grew steadily stronger toward the end of the twentieth century, it would be hard to know whether the boundaries of genres were actually strengthening, or just aligning more closely with 1990s-era expectations.

Our second experiment tried to avoid this source of anachronism by basing genre classifications on Kirkus reviews contemporary with the works of fiction themselves. One limitation here is that we only have Kirkus reviews after 1928 (and have very few of them until the early 1930s). This strategy also still leaves some ambiguity about the exact chronological perspective represented. While reviews were usually written within a few years of a book's publication, our topic model of reviews spans a much longer period, and the boundaries of topics might be shaped by patterns across the whole timeline. Also, we selected specific topics as "genre-like" and ignored others (since evaluative discourse, for instance, is important in reviews, but no one thinks "bad writing" is a genre). That activity of selection could conceivably add a contemporary bias. To minimize the risk, we did not group topics. For instance, we identified several different topics that loosely aligned with our concept of "detective fiction," but we allowed each of those topics to count as a distinct genre instead of using our judgment to fold them together.

3. Methods

3.1. General considerations and limitations

The best way to measure the differentiation between literary genres is probably to train supervised predictive models that attempt to distinguish works in one genre from other works in a given period or cultural milieu. It may seem odd to measure differentiation textually, since we have stressed that we understand genres as perspectival social categories–not necessarily defined by any fixed textual template. But it nevertheless turns out that genres can be discriminated quite reliably using textual features. Moreover, the textual signature of a genre tends to be stronger when human beings agree strongly about the boundaries of the category [4, 6]. So the textual coherence of genre categories can serve as a rough proxy for the coherence of social conventions and expectations governing readers and writers who are no longer alive to be interviewed.

Supervised models are a good way of measuring textual coherence because they don't require a researcher to make risky assumptions about the specific textual patterns that define genres. Should we emphasize common function words? content words? It's hard to know *a priori*. But a supervised model can assign weight to the features that do in practice distinguish two categories. (In practice, both function words and content words are significant.) When a supervised strategy has been tested against others, human judgment correlates more closely with the accuracy of supervised models than with distance measurements based on word counts or topic proportions [18].

However, supervised modeling is also difficult to apply across a long timeline, because we need a reasonable number of works to train a model. If our collection contains only two "imaginary voyages" in the 1890s, it will be very difficult to define a meaningful model of the genre. This problem is substantial, because real collections are often quite sparse in early periods.

An alternate approach is simply to measure the distances between texts. If genres are closely-knit and crisply differentiated, for instance, we would expect the distance between two books in the same genre to be lower, on average, than the distance between a pair of randomly selected books. This strategy has the advantage that it can apply even to sparsely represented genres. As long as a given period includes at least two books in genre G, the distance between them can be measured. And, while this metric is not quite as flexible as a supervised strategy, earlier work has found that it does in practice correlate closely with the results of supervised modeling [6]

This simple comparison between distances loosely resembles several systems of measurement that have been applied to validate clustering algorithms [11]. But in validating clustering, researchers can make additional assumptions that don't apply well to genre. For instance, in regular k-means clustering, each point belongs to one and only one cluster. It therefore makes sense to assume that clusters should be mutually exclusive. Instead of subtracting intra-cluster distances from distances in the dataset at large, a metric like "silhouette value" accordingly subtracts intra-cluster distances from the average distance to the next nearest cluster—a tactic that will tend to penalize partitioning schemes with clusters that overlap.

It would be inappropriate to apply this metric to genre metadata, because a book can belong fully to several genres at once. A mystery can also be, for instance, a historical novel, and a historical novel can be a love story. Libraries record many examples of this kind, and it is not intuitively obvious that the categories involved are weakened by these areas of overlap. Just as importantly, many works of fiction were never assigned to any subgenre at all. So we probably don't want a metric that assumes an exhaustive partitioning of the data.

3.2. The specific method we used

For all these reasons, it seemed best to adopt a very simple measurement strategy, based on contrasting pairs of books that share a genre tag to pairs of books that don't. Random subsampling also seemed appropriate, because the total number of books is in the thousands and exhaustive permutation would be slow.

We want to compare the strength of genre groupings in different periods, so we will compare pairs of books that were published near the same date. If we had enough books, we could insist on comparing books published in the same year. But since some genres are sparsely represented, we compare pairs published within ten years of each other, to ensure books in uncommon genres can find a match. Because it is conceivable that the exact number of years separating a pair of books could matter, and could vary across the timeline, we ensure that volumes selected for out-of-genre comparisons are precisely matched to the dates of the in-genre pair.

We used the 2500 most frequent words in our sample, and standardized the document-feature matrix using the same standard scaling procedure we would use for predictive modeling (that is, we transformed each column into a z score by subtracting the mean and dividing by standard deviation). This standardization process has been widely adopted, not only in machine learning, but in stylometry where it goes under the name of Burrows's Delta. Cosine distances on Delta-normalized feature vectors have outperformed other strategies in authorship attribution, and feature counts in the range of 1000 to 5000 words have proved effective both there and in genre classification [8, 2].

In pseudo-code, the general form of the algorithm is this:

```
for an arbitrary number of iterations (40,000):

    randomly select a book g1 from the list of books with genre tags;
    if it has multiple genre tags, randomly select one as G

    find another book g2 in genre G, and published within 10 years of the
    publication of g1, but not by the same author

    find a book o1 published in the same year as g1, but not sharing any genre
    tags with g2 (and not by the same author)

    find a book o2 from the same year as g2, but not sharing any genre tags
    with g1 (and not by the same author)

    measure cosine distance between g1 and g2; this is the in-genre distance

    measure cosine distances between g1 and o2, and between g2 and o1;
    the mean of these two measurements is the out-of-genre distance

    from out-of-genre distance, subtract the in-genre distance;
    associate the difference with the mean of the two publication dates
```

Broadly speaking, this algorithm is a "matching method"; it tries to find pairs of books that are similar except for the fact that one pair shares a genre tag and the other pair(s) don't. A few details in the matching strategy are worthy of notice. Each time we select a book we randomly select one of its genre tags. We don't iterate through all the tags: if we did, books with multiple genre tags would be overrepresented in our sample. In selecting out-of-genre comparisons we ensure that the chronological distance between books is always the same as it was for the in-genre comparison. Finally, we avoid comparing books by the same author,

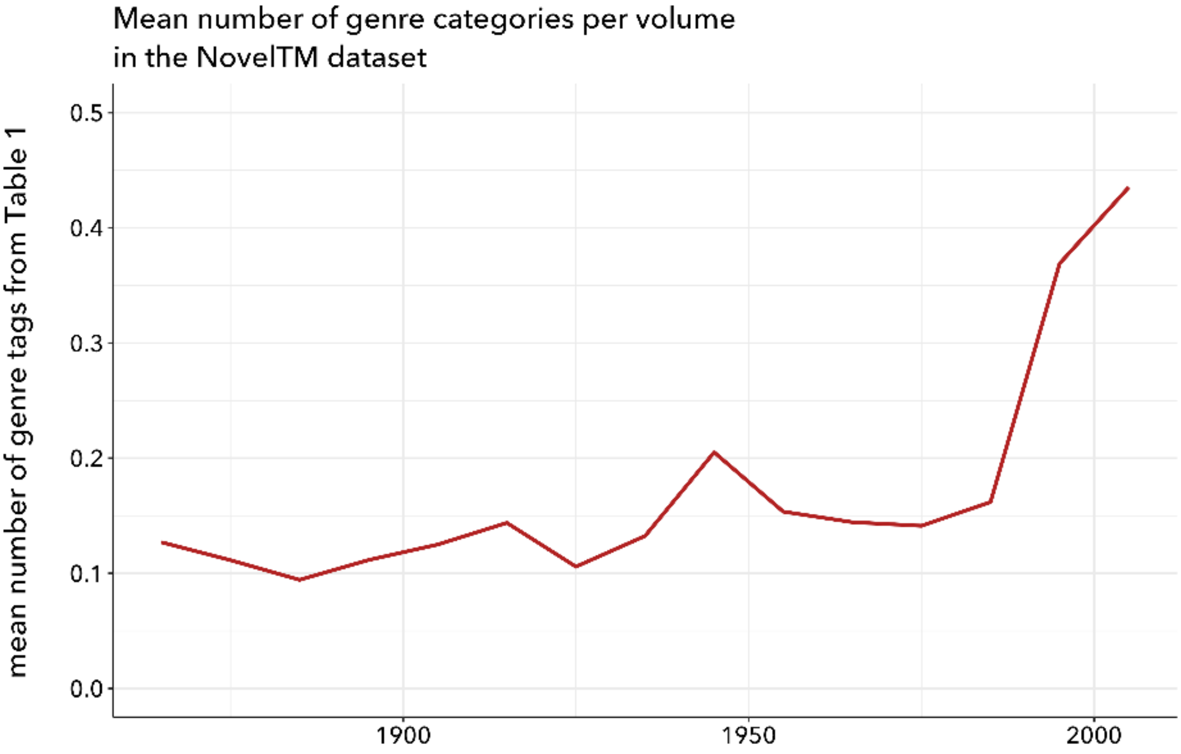

Figure 1: The mean number of genre categories (among those included in our experiment) assigned to a volume in a given decade. Most volumes are uncategorized.

because those distances are likely to be much lower, and it is possible that the probability of a same-author pairing could be unevenly distributed across the timeline.

This experiment was repeated for Library of Congress tags assigned to books by catalogers (1860-2009), and for genre-like topics inferred from *Kirkus Reviews* (1928-2009).

We also preregistered a plan to run this experiment in two different forms. Since many works of fiction are not associated with any genre category at all, selecting out-of-genre contrast volumes (*o1* and *o2*) only from other genres included in our experiment might not give us a broad representation of fiction at large. Notice, in Figure 1, that genre tags are actually quite rare until the end of the twentieth century. Even then, less than half the volumes bear a label from Table 1.

We therefore used two different contrast sets: one drawn randomly from the whole NovelTM fiction collection (or, in evaluating Kirkus topics, from the part of the collection reviewed in *Kirkus Reviews*), and one drawn only from the subset of books that also carried at least one genre tag.

It wasn't immediately obvious to us which of those methods should be preferred. The "fully random" contrast set provides a better measure of the way genres are differentiated from fiction at large. But a contrast set drawn from "other genres" would be more informative if researchers were interested in the *internal* differentiation of the genre system. So we preregistered both experimental plans. As it turns out, they produce strongly parallel results.

Table 1
The nineteen categories used in our Library of Congress experiment.

historical fiction	legends	fairy tales	folklore
domestic fiction	personal narratives	adventure stories	psychological fiction
war stories	detective fiction	western stories	imaginary voyages
sea stories	ghost stories	short stories	tales
humor	love stories	science fiction	

3.3. Genre-like categories

Our goal was to measure the differentiation of fiction across a long timeline from 1860 to 2009. In the early part of that period, one could argue that fiction is not differentiated by genre at all, but by subject. However, some of those subject categories later turned into genres, and the gradual character of the transformation is precisely what historians need to understand. So it wouldn't have made sense for us to draw a sharp boundary between genres and subjects. Excluding either side of that boundary would exclude exactly the gradient we are investigating.

We therefore considered Library of Congress subject headings as well as genre/form headings in our experiment. Many of the categories we used (like "science fiction" and "fairy tales") will readily be called genres by most readers; others (like "sea stories" or "war stories") may be closer to subjects. But the distinction is blurry. All nineteen categories are listed in Table 1; note that several different strings often mapped to each category. (E.g. "shipwrecks" and "castaways" mapped to "sea stories.")

We also included "short stories," although the short story is arguably a form rather than a genre. This was debatable, but we found it better to err on the side of inclusion. The category "personal narratives" includes narratives that may be based on the author's own life, but that seem formally closer to fiction than biography. The topics selected as "genre-like categories" in our topic model of *Kirkus Reviews* are described (through keywords and examples) in supporting material online, along with the rest of our data and code [21].

3.4. Preregistered hypotheses

We preregistered three hypotheses for this experiment [20].

3.4.1. Hypothesis 1

For the experiment using Library of Congress Headings, we expect that there is some year Y in the second half of the twentieth century, such that the average generic closeness (out-genre-distance minus in-genre-distance) increases from 1860 to that year, and decreases from that year to 2009, where "increases" and "decreases" describe a correlation with time. We expect both correlations to be significant at $p < .05$.

3.4.2. Hypothesis 2

For the experiment using a topic model of *Kirkus Reviews*, we expect the average out-topic/in-topic difference, aggregated by year, will correlate with the yearly measurement in our first experiment in the same time period (1930–2009).

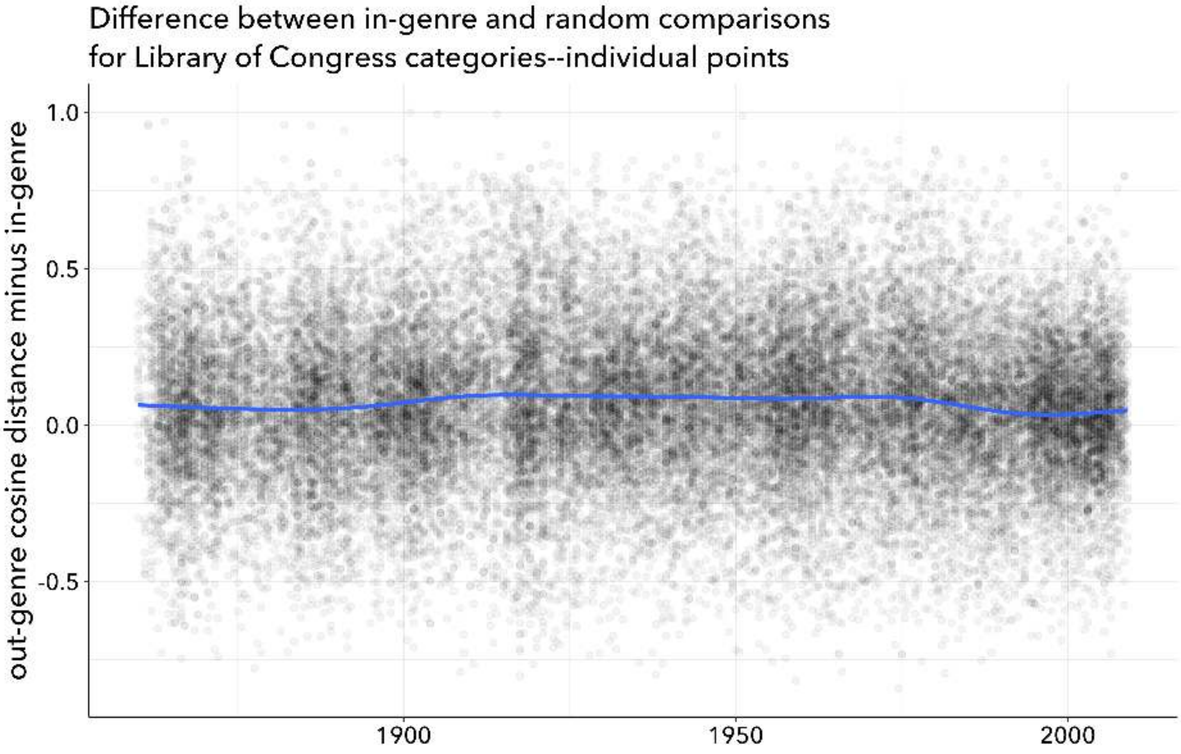

Figure 2: Each point marks the difference between a) the cosine distance between two volumes in the same genre and b) the cosine distance between those volumes and others not in the same genre.

3.4.3. Hypothesis 3

We further speculated that the out-topic/in-topic difference would be strongest in decades when a genre was most prevalent in *Kirkus Reviews*. In other words, science fiction would be more clearly differentiated from other genres in decades when science fiction was common.

4. Results

Measuring cosine distances between individual pairs of books produces results that vary enormously from one pair to the next. Fig. 2, for instance, shows the raw differences between in-genre and out-genre cosine distances in one experiment. (We use Fisher's z-transform on the cosines before subtracting them, since a difference of .01 between two cosines can correspond to a different angular magnitude depending on the absolute value of the cosines.)

The trend line in Fig. 2 does bend significantly. But the change in the trend is dwarfed by the variation between individual pairs of books. This doesn't necessarily mean that the change is objectively small. When researchers use predictive models to study the cohesion of genres, they find changes of significant magnitude across this period. Science fiction, for instance, can only be identified with 75% accuracy in the late nineteenth century, but approaches 95% in the middle of the twentieth [17]. The noise in Fig. 2 is better understood as a consequence of methodology: in order to cast a broad net across many genres at once, we chose to use textual distances rather than predictive models. And the distance between a single pair of volumes is

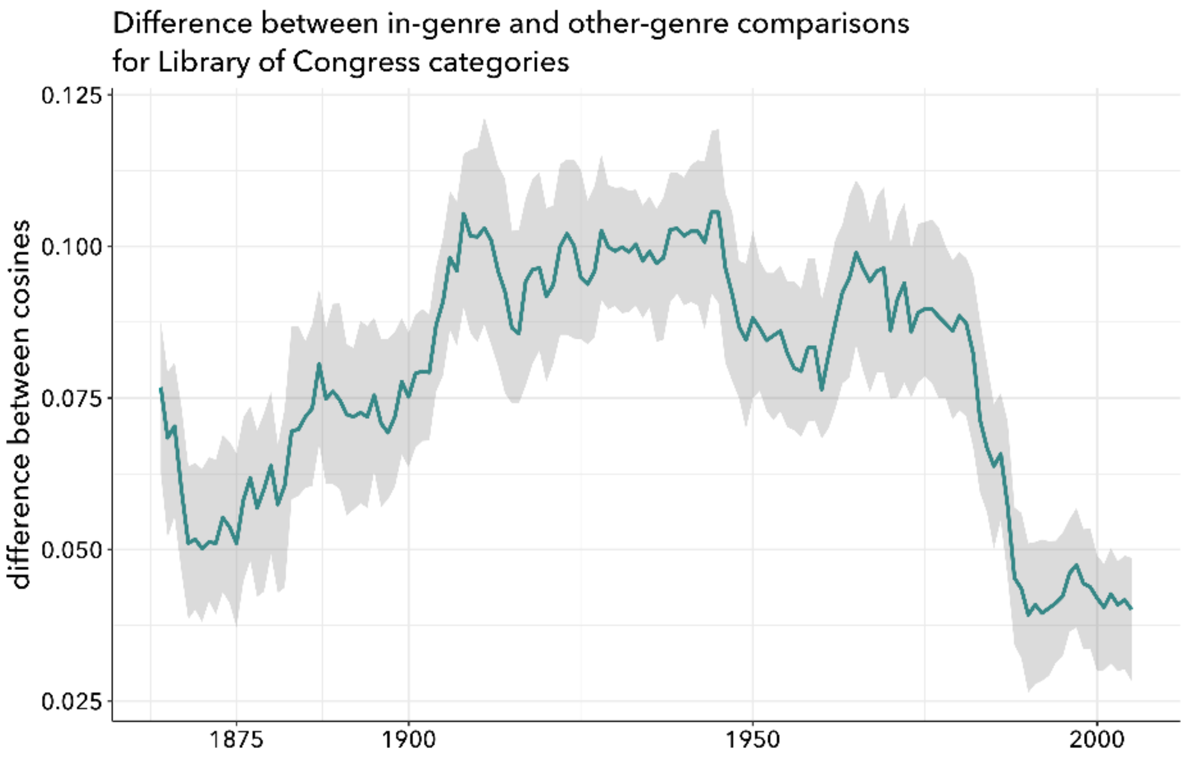

Figure 3: Running median of the difference between in-genre distances and distances to volumes in a different genre. The median of a nine-year window is plotted at the midpoint for the window, along with a shaded 95% confidence interval for the median, calculated by bootstrap resampling.

an inherently variable and noisy measurement, affected by many factors other than genre.

There are, however, significant changes over time, confirming most of our preregistered hypotheses. To illuminate those changes we plot results below using a running median in a nine-year moving window.

4.1. Experiment using Library of Congress headings.

Our first experiment was conducted using 19 genre-like categories based on headings defined by the Library of Congress and assigned to volumes by librarians (usually retrospectively). See Table 1 for a full list.

We compared in-genre similarities to two different contrast sets. First, comparisons between volumes in different genres, in Fig. 3.

The strength of generic clustering peaks toward the first half of the twentieth century, not the second half, as we had expected. But our hypothesis was defined loosely enough that it can be confirmed by this data. We predicted there would be a year Y in the second half of the twentieth century with a statistically significant positive trend up to Y, and a statistically significant declining trend thereafter. We didn't insist on identifying the exact moment when the slope of the trend changes (which, in this case, might be around 1942).

We also ran this experiment using a sample of fiction randomly selected from HathiTrust (including volumes not assigned to a genre at all). The results are visualized in Fig. 4, and

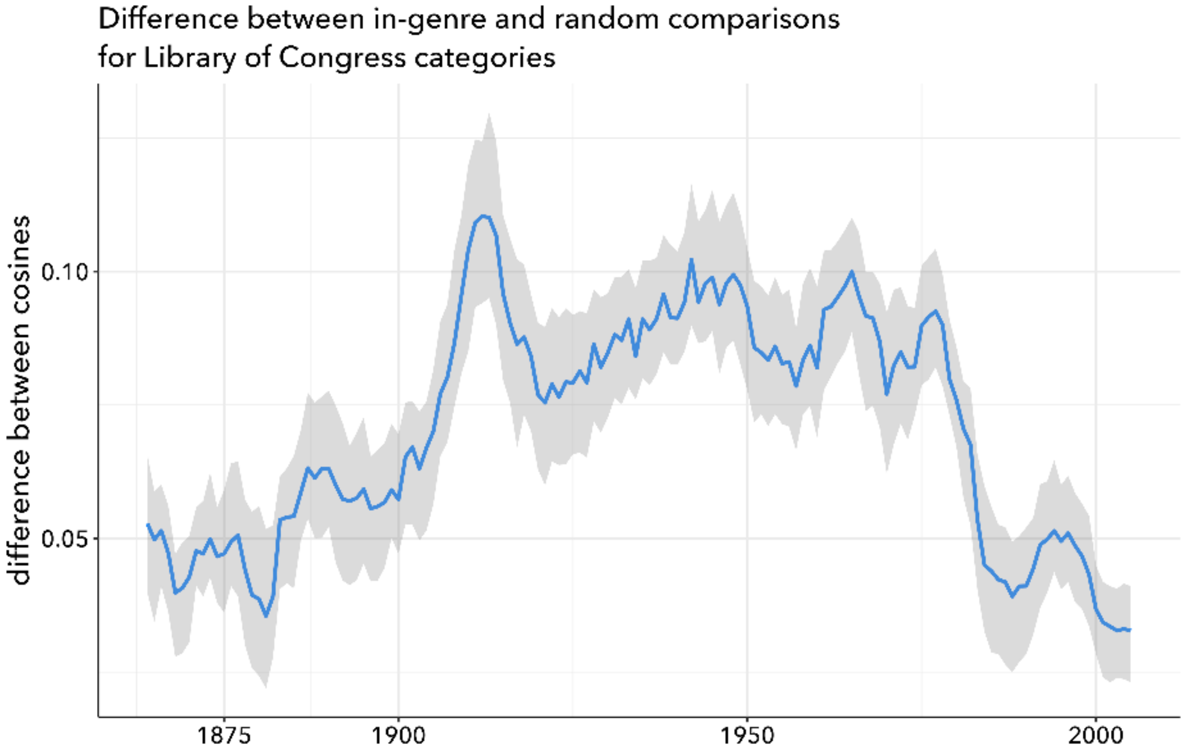

Figure 4: Running median of the difference between in-genre distances and distances to volumes of fiction selected randomly from the whole NovelTM dataset. The median of a nine-year window is plotted at the midpoint for the window, along with a shaded 95% confidence interval for the median, calculated by bootstrap resampling.

the trend is broadly similar. Again it appears to peak around the middle of the twentieth century. It is difficult to know whether smaller features (like the brief apparent peak around 1910) represent interpretable changes in literary history or just sampling noise.

Several problems we had anticipated don't actually appear in this data. For instance, we had worried that we might see a steady rising trend all the way to the end of the timeline. If that had happened, it would be possible to wonder whether our categories (mostly applied by observers since the last few decades of the twentieth century) are simply a better fit for volumes in that period. But that is clearly not a problem we confront here; the decline in generic cohesion since 1980 is in fact quite dramatic.

But there are still reasons to wonder whether the patterns seen above may be an artefact of the history of library classification. For instance, as Figure 1 makes plain, the proportion of volumes tagged with a genre label is much higher in the last twenty years. It seems possible that this could have the reverse of the distorting effect we anticipated: it could dilute and blur genre categories. Perhaps in earlier years only the most obvious examples of a genre got tagged, but after 1990 or so tags were applied to borderline cases? That could explain the sudden drop at the end of both trend lines above. To address questions like this, we tried to confirm this trend using book reviews.

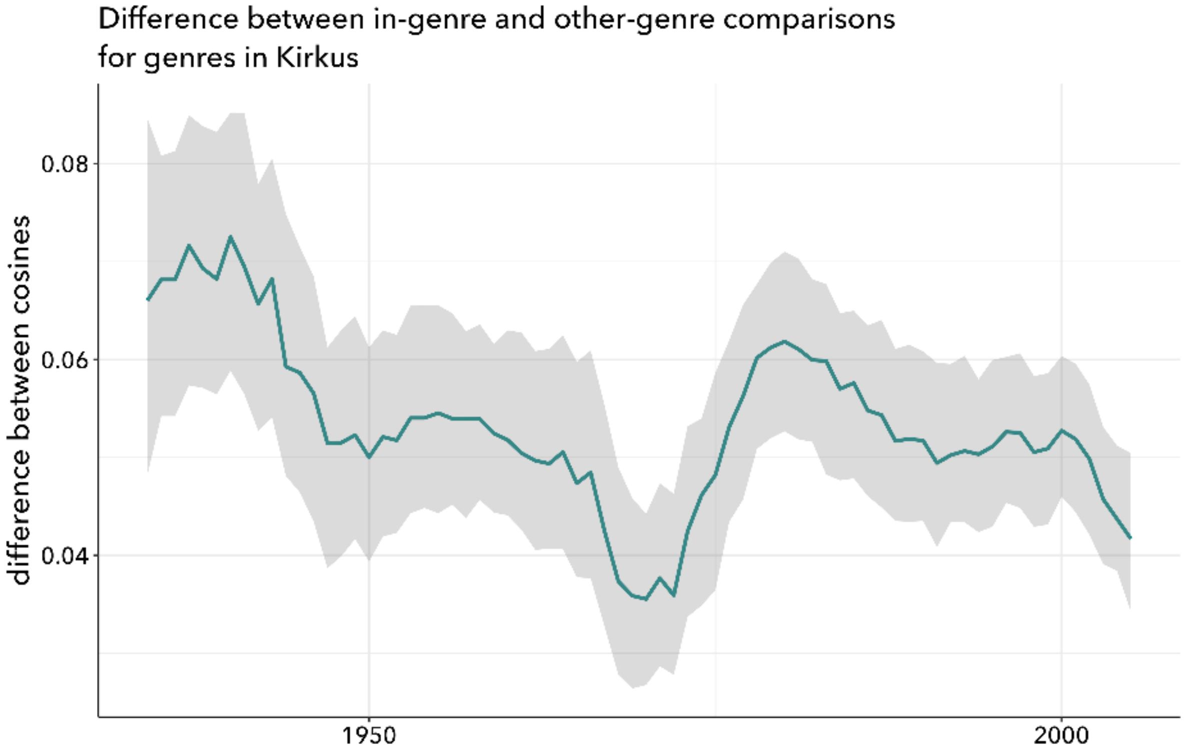

Figure 5: Differences between in-genre and other-genre comparisons for books whose reviews in *Kirkus Reviews* were associated with 23 genre-like topics. We plot the running median of a nine-year moving window.

4.2. Experiment using a topic model of *Kirkus Reviews*.

Kirkus Reviews is a long-running book review magazine founded by Virginia Kirkus in 1933. The online reviews we gathered presumably include some reviews she wrote earlier, while working for Harper & Brothers in the 1920s. Most of the reviews are short—rarely more than a couple of paragraphs—and rapidly get to the point where genre is concerned. Here is the entire review of Raymond Chandler's hard-boiled detective novel *The Big Sleep* (1938), for instance: "A good one in the tough school, in which private detective Marlowe is hired to investigate a blackmailing and finds himself bucking a well-run gang, several murders, and the DA's office. Hard-boiled, fast paced, plenty of action, some sensationalism. Not for conservatives" [10].

Reviews like this signal genre fairly clearly: the words "hard-boiled" and "detective," for instance, are explicitly present in that review of *The Big Sleep*. It therefore seems to make sense to use a topic model of the reviews as evidence of genre categorization. We can attempt to measure generic cohesion by measuring the similarity of book texts grouped by reviewers. This method has the advantage of using testimony contemporary with the books themselves. But contemporaneity can be double-edged. In periods where genres are still too inchoate to be recognized by reviewers, this method might not reveal very much about them. It can reveal the waxing or waning of genre differentiation only after a genre has crystallized enough to implicitly organize review discourse. A skeptic might even worry that this perspective tells us less about the absolute coherence of a genre than it does about the difference between its coherence in

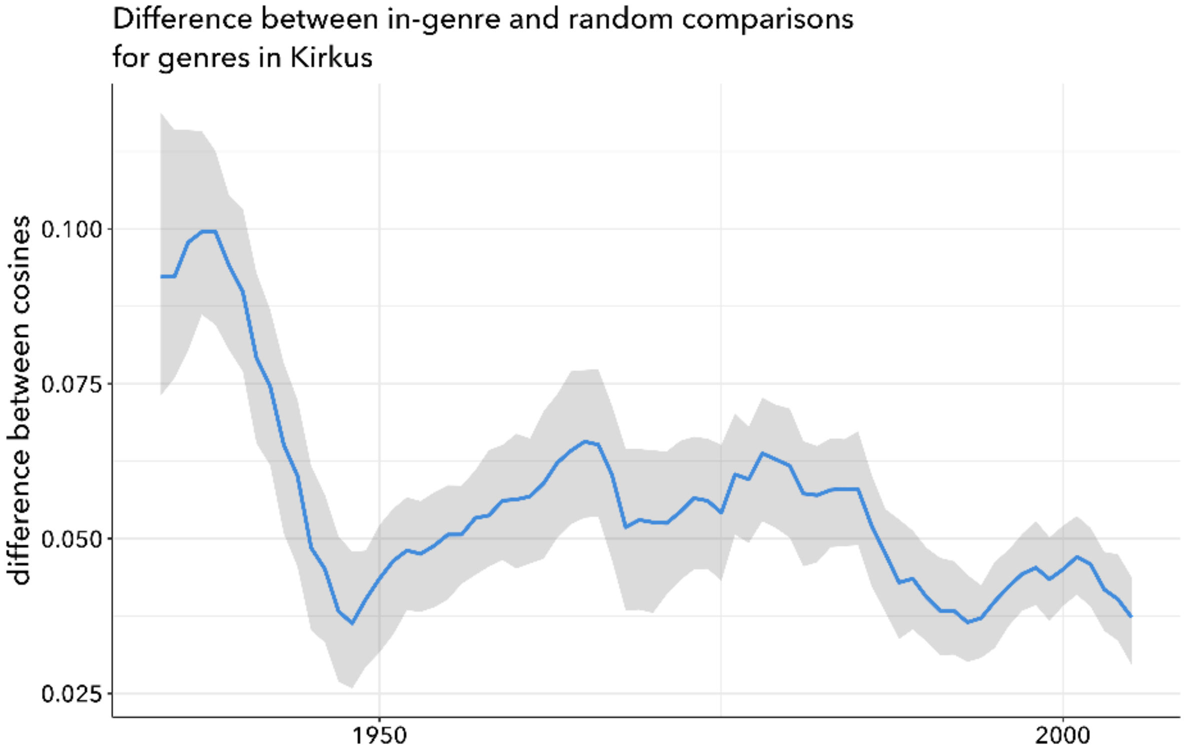

Figure 6: Differences between in-genre and random out-of-genre comparisons for books whose reviews in *Kirkus Reviews* were associated with 23 genre-like topics. We plot the running median of a nine-year moving window.

reviews and in the books themselves. It is, however, *different* from the retrospective vantage point of librarians, and a difference of perspective was what we chiefly needed.

We topic modeled 19,018 reviews from Kirkus. First, we used the topic coherence measure implemented in Gensim to optimize the number of topics for this particular dataset [16]. Searching across a range from 10 to 90 topics, we found topic coherence optimized at 80. Therefore, we set the number of topics to 80 and implemented Mallet's LDA algorithm with Gensim [14].

We identified 23 topics as genre-like categories. These were not always precise matches for any specific Library of Congress genre; for instance, one topic with keywords "case, murder, lawyer, trial, charge, wife, kill, law, evidence, accuse," could have been called either "detective fiction" or "legal thriller." But whatever the label, it seemed suggestive of a genre. We also found topics suggestive of domestic fiction, romance, historical fiction, comedy, erotic fiction, ghost stories, and science fiction. Of the 19,018 reviews we had, 7,133 showed one of these genre-like topics as the most prominent topic. Note that unlike the LoC categories, these are mutually exclusive categories, since a review can only have one most prominent topic.

Treating each topic as a "genre," we measured the difference between in-genre and out-genre distances (as in the previous experiment) for all volumes whose reviews were dominated by one of the 23 genre-like topics. We found patterns that loosely match the trend in our Library of Congress experiment, in the sense that there is a decline from 1934 to 2005. (Note that Kirkus doesn't cover the first half of the timeline in our earlier experiment, from 1864 to 1933.)

We preregistered a hypothesis that there would be a significant correlation between the trend lines here and in our earlier experiment. That turns out to be true. For the other-genre contrast plotted in Fig. 5, the correlation is $r = .30, p < .01$, if we treat each year's median as a separate observation.

For the fully random contrast in Fig. 6, the correlation is $r = .37, p < .01$. It's not a very strong correlation in either case; the specific contours of the trend are different. For instance, in the Kirkus dataset, much of the decline takes place early (by 1950), whereas in the Library of Congress dataset, genres become blurry mostly in the last 20 or 25 years.

However, while this signal is weak, there is some reason to believe it real. For instance, we have noticed what medical researchers call a dose-response relationship: the correlation between the LoC and Kirkus trend lines appears to get stronger as we focus the Kirkus datasets on books more clearly assigned to a specific genre. If we include all 80 topics in the experiment (not just the 23 labeled as genres), the correlation goes away entirely. On the other hand, if we restrict our genre corpus to 3,810 books where those 23 topics were very prominent (above the median for the most-prominent-topic in a review), we get correlations of $r = .51$ to .53 instead of .30 to .37. So we feel our second preregistered hypothesis is also solidly confirmed.

Our third preregistered hypothesis was that, in general, genres would be most strongly differentiated when a genre was most prominent in the corpus. We tried to confirm this hypothesis by measuring the proportion of volumes assigned to each genre (i.e., topic) in each decade, and correlating it with the mean in-genre/out-of-genre difference for each genre in each decade. But in our Kirkus data we see no relationship at all; r is close to 0, and it is impossible to reject the null hypothesis.

5. Interpretation

There are several reasons to interpret these results cautiously. We have already mentioned that genre is a perspectival concept. Different observers define it differently. We have compared two different sources above, but many others could be consulted.

More fundamentally, the trend traced above will only seem important and meaningful to readers who believe genres are likely to evolve in parallel, as a "genre system." There is some consensus that this is true in the first half of the timeline we study. John Rieder, at any rate, has argued that science fiction, the western, the detective story, and so on, need to be understood collectively as part of a "mass cultural genre system" driven by the rise of commercial advertising in the late nineteenth and early twentieth centuries [15, p. 33–64].

But there is no equivalent consensus about the decline of the genre system in the second half of the twentieth century. We believe the results of our experiment suggest that there was such a decline, equal in magnitude to the consolidation of genre at the beginning of the century. But this thesis will probably only be accepted after literary historians closely investigate the histories of individual genres to ascertain that they did, in fact, move in parallel. The methods we have used above are not sensitive enough to distinguish those stories.

We have considered and weighed some obvious quantitative objections to the premise that a "genre system" waxed and waned. One concern is based on Simpson's paradox: our story about a generalized rise and fall of genre differentiation might really be a story about the rise and fall of a small subset of strongly differentiated genres. The sample we have described above is open to that objection because it allows the sizes of genres to vary in proportion to their prevalence in academic libraries. So if, say, detective fiction was always strongly differentiated,

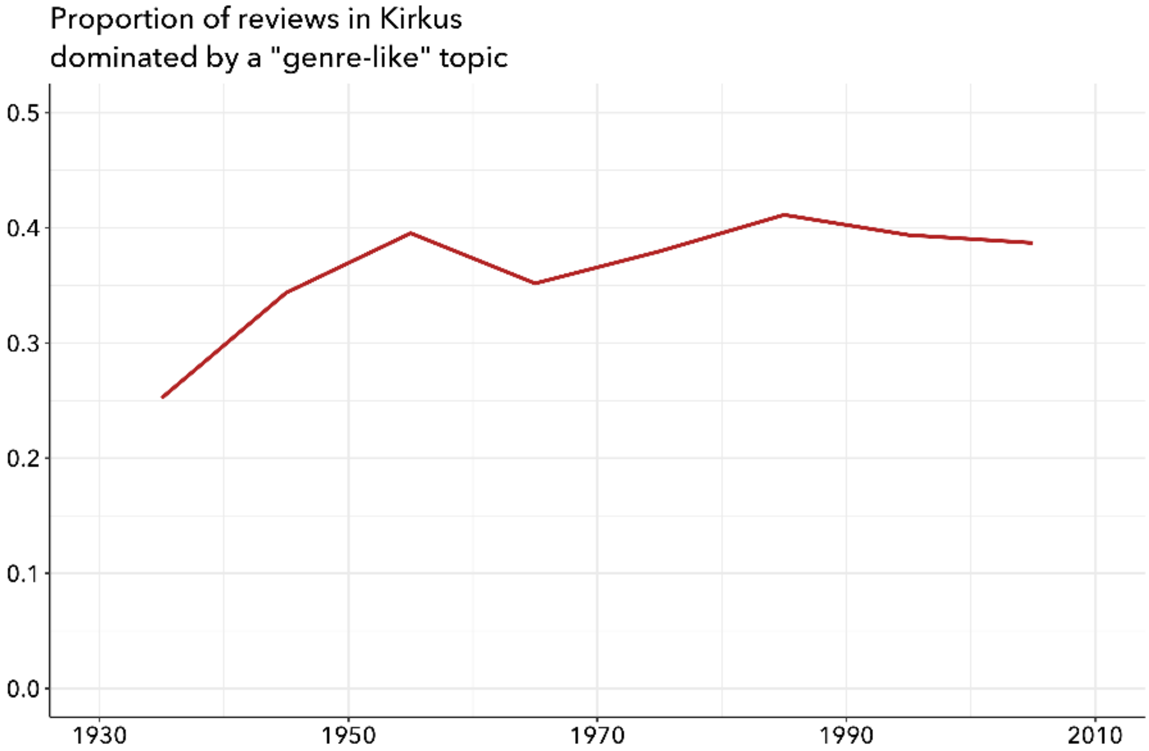

Figure 7: The proportion of books in *Kirkus Reviews* whose reviews are dominated by one of our 23 "genre-like" categories.

and the number of detective stories greatly expanded in the middle of the twentieth century, the differentiation of genres collectively might appear to rise and fall (because the population was changing) while the level of differentiation in individual genres was really entirely stable.

We ran several tests to detect a pattern of this kind, and while we don't have room here to report results in depth, the trend we have described above seems to be robust to all the sampling strategies we have so far tried. For instance, we also ran our Library of Congress experiments in a sample where the sizes of genres were capped, to keep genre distribution as constant as possible across the timeline. We still saw a significant rise and fall of differentiation in that sample.

However, there are some fundamental interpretive paradoxes in this topic that we cannot resolve by resampling or by running statistical tests. For both of the sets of observers we considered, the late-twentieth-century decline of genre differentiation closely coincided with a substantial increase in the proportion of fiction that observers recognized as belonging to a genre.

In Fig. 1, we pointed out that the proportion of fiction librarians tagged with genre categories shot upward at the very end of the twentieth century. A rising trend is also perceptible in *Kirkus Reviews*, although the rise is weaker and takes place earlier.

This could be viewed simply as a source of distortion, if we assume that observers' criteria for placing a book in a genre loosened as time passed. In that case, genre boundaries might appear to get blurrier in the late twentieth century merely because more borderline cases were getting a genre label.

But the growing number of books with genre labels also poses other interpretive puzzles. Even if genre boundaries really did become blurrier, we're faced with the paradox that the genre system was simultaneously expanding, and categories were subdividing. It's not impossible for all three of those things to happen at once. Specialized forms of science fiction like "cyberpunk" and "steampunk" could proliferate, for instance, even while science-fictional themes were leaking into other genres and the boundaries of SF itself were growing more diffuse. But a tangled situation of that kind would need to be described very carefully. We probably wouldn't say, for instance, that the genre system as a whole was growing less important—even if it was, at the top level, becoming less clearly differentiated.

In short, 150 years of literary history constitute a large, messy object that will need to be studied from many different angles. This is just a first pass at a quantitative description of genre differentiation, so we have tried to present our conclusions modestly. But all the evidence we have seen so far is consistent with an account that shows genre differentiation rising to (roughly) the middle of the twentieth century, and declining at some point before the century's end.

6. Contribution statement

All authors contributed to framing research questions, interpreting results, and writing the final report. In addition:

Aniruddha Sharma extracted an additional dataset from *Publishers Weekly,* and performed text clustering on the same that shaped our interpretation of results, although the data is not directly included above.

Yuerong Hu performed topic modeling on the extracted *Kirkus* book reviews; her model was the foundation for all our experiments on *Kirkus.*

Peizhen Wu categorized topics as genre-like and did research on the history of genre.

Wenyi Shang developed an additional dataset and performed intricate pattern matching techniques that shaped our interpretation of results, although that data is not directly included above.

Shubhangi Singhal scraped all data from *Kirkus Reviews* and deduplicated it, correcting missing and erroneous entries.

Ted Underwood obtained funding, produced visualizations for the paper, and was the corresponding co-author.

Author order was determined randomly, with the proviso that Underwood went last.

Acknowledgments

Material support for this project came from NovelTM, directed by Andrew Piper and funded by Canada's Social Sciences and Humanities Research Council.

References

[1] C. A. P. Allen. "The Informed Victorian Reader". PhD thesis. Ann Arbor, MI, USA: University of Michigan, 2016.

[2] S. Argamon. “Interpreting Burrows’s Delta: Geometric and Probabilistic Foundations”. In: *Literary and Linguistic Computing* 22.2 (2008).

[3] D. Biber and S. Conrad. *Register, Genre, and Style*. Cambridge, UK: Cambridge University Press, 2019.

[4] J. Calvo Tello. *Genre Classification in Spanish Novels: A Hard Task for Humans and Machines?* 82. 2018. URL: https://eadh2018.exordo.com/programme/presentation/82.

[5] B. Capitanu et al. *The HathiTrust Research Center Extracted Feature Dataset (1.0)*. 2016. DOI: 10.13012/J8X63JT3.

[6] K. Chang et al. *Book Reviews and the Consolidation of Genre*. 2020. DOI: 10.17613/02q2-1v27.

[7] R. Cohen. “History and Genre”. In: *New Literary History* 17.2 (1986), pp. 203–218.

[8] S. Evert et al. “Understanding and Explaining Delta Measures for Authorship Attribution”. In: *Digital Scholarship in the Humanities* 31.2 (Dec. 2017).

[9] F. Jameson. *Postmodernism, or, the Cultural Logic of Late Capitalism*. Durham, NC, USA: Duke University Press, 1991.

[10] V. Kirkus. “Review of The Big Sleep”. In: *Kirkus Reviews* (Feb. 1938).

[11] Y. Liu et al. “Understanding of Internal Clustering Validation Measures”. In: *2010 IEEE International Conference on Data Mining*. Sydney, NSW, 2010, pp. 911–916.

[12] D. P. Miller. “Out from Under: Form/Genre Access in LCSH”. In: *Cataloging and Classification Quarterly* 29 (2000), pp. 169–188.

[13] S. Rachman. “Poe and the Origins of Detective Fiction”. In: *The Cambridge Companion to American Crime Fiction*. Ed. by C. R. Nickerson. New York, NY, USA: Cambridge University Press, 2020, pp. 17–28.

[14] R. Řehůřek and P. Sojka. “Software Framework for Topic Modelling with Large Corpora”. English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[15] J. Rieder. *Science Fiction and the Mass Cultural Genre System*. Middletown, CT, USA: Wesleyan University Press, 2017.

[16] M. Röder, A. Both, and A. Hinneburg. “Exploring the Space of Topic Coherence Measures”. In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. WSDM ’15. Shanghai, China: Association for Computing Machinery, 2015, pp. 399–408. ISBN: 9781450333177. DOI: 10.1145/2684822.2685324. URL: https://doi.org/10.1145/2684822.2685324.

[17] T. Underwood. “Machine Learning and Human Perspective”. In: *PMLA* 135.1 (Jan. 2020), pp. 92–109.

[18] T. Underwood. “The Historical Significance of Textual Distances”. In: *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Santa Fe, New Mexico: Association for Computational Linguistics, Aug. 2018, pp. 60–69. URL: https://www.aclweb.org/anthology/W18-4507.

[19] T. Underwood, P. Kimutis, and J. Witte. “NovelTM Datasets for English-Language Fiction, 1700-2009”. In: *Journal of Cultural Analytics* (May 2020). DOI: 10.22148/001c.13147.

[20] T. Underwood et al. *Preregistration for Book Reviews and the Consolidation of Genre*. Sept. 2020. DOI: 10.17605/OSF.IO/MDVTB. URL: osf.io/mdvtb.

[21] T. Underwood et al. *tedunderwood/riseandfall: Code repo for The Rise and Fall of Genre Differentiation*. Version 1.1. Sept. 2020. DOI: 10.5281/zenodo.4041115. URL: https://doi.org/10.5281/zenodo.4041115.

[22] G. K. Wolfe. *Evaporating Genres: Essays on Fantastic Literature*. Middletown, CT: Wesleyan University Press, 2011.