sociological science

Supplement to:

Underwood, Ted, Kevin Kiley, Wenyi Shang, and Stephen Vaisey. 2022. "Cohort Succession Explains Most Change in Literary Culture." Sociological Science 9: 184-205.

SM 1 Data and topic model

SM 1.1 Sources of date information

The large collection of fiction we borrowed from Underwood et al. (2020) provides an "estimated latest possible date of composition" instead of a confirmed date of first publication. This estimate is the earliest date attested for a title in HathiTrust Digital Library, or the author's year of death (when that is earlier than the earliest library copy).

To create a more accurate estimate of publication date, we collated this collection with dates manually assigned at the Chicago Text Lab (by Hoyt Long, Edwin Roland, and Richard Jean So) and at the University of Illinois (by Patrick Kimutis, Wenyi Shang, Ted Underwood, and Jessica Witte). We also used the United States Copyright Registry to correct dates that erred on the late side. Since we only had a partial sequence of the registry starting in 1923, and there can be multiple registrations for a title, we couldn't assume that the registry always provided the correct date. But when a book appears earlier in the registry than in library metadata, we could assume that the earlier date was closer to being correct. We also continued the practice of capping publication date at the author's year of death, since the author is unlikely to have made substantial changes afterward.

This process gave us a sample of 10,830 works that we actually used in analysis, divided into two overlapping subsets.

For the regression experiment, we felt it was advisable to hold the national composition of the dataset constant across the timeline to avoid spurious effects. So, for that experiment, we could only use 5,572 works where we had confirmed the author resided in the US for a significant portion of their career. If the full metadata `corpus` is a Pandas dataframe, the subset used in regression can be selected using the following conditions:

```
regression_set = corpus.loc[(corpus.pubdate_known == True) &
```

```
(~pd.isnull(corpus.birthyear)) &
(corpus.firstpub >= 1890) &
(corpus.firstpub <= 1989) &
(corpus.us_national == True), : ]
```

In structural equation modeling, results depended only on longitudinal comparisons *within* an author's career—not between authors—so it was possible to use authors of unconfirmed nationality. (Given the composition of the data, most unconfirmed authors are still in practice authors of US nationality.) However, we could only use authors who had at least three works with confirmed publication dates in our corpus. This gave us 10,355 works that can be selected with the following command:

```
sem_set = corpus.loc[(corpus.pubdate_known == True) &
(~pd.isnull(corpus.birthyear)) &
(corpus.firstpub >= 1880) &
(corpus.firstpub <= 1999) &
(corpus.authof3ormore == True), : ]
```

Note that the date limits are slightly more inclusive here, for reasons explained below.

SM 1.2 Distribution of works across time

The complete set of works in the topic model is selected to produce an even distribution of words across time from 1880 to 1999. This keeps the "granularity" of topics roughly comparable across the timeline. If the first half of the timeline had many fewer words than the second half, there would tend to be fewer topics there, and styles specific to that period would be divided more coarsely.

Keeping granularity constant may be an unnecessary precaution: there's not actually any clear reason to believe that variation would be problematic for the questions we pose in this paper. For the most part, our regression and SEM experiments target individual topics, and differences of granularity *between* topics should have no effect on the experiment. But it is conceivable that variation would create a spurious "period effect," since it might cause topics to increase in prominence in areas of the timeline where there are fewer competitors. This could, at least theoretically, distort the regression experiment. In any event, we wanted to model a relatively large corpus, and it was easy to select books to produce a uniform distribution.

Through a similar abundance of caution, we extended the corpus ten years on either side of the century (1890–1989) we had decided to primarily target, because granularity can tend to decline toward the edge of a topic model. Since the SEM experiment seemed less likely to be sensitive to this variation, we did allow a few triplets in that part of the experiment to extend into the 1880s or 1990s.

In figure SM 1 we visualize the distribution of both corpora across the two timelines relevant to this study: publication years and birth years. The corpus

used for longitudinal study (SEM and distance measurement) is more sharply concentrated in the middle of the 20th century, because it relies more heavily on the copyright registry. The birth year distributions cover a slightly wider span than publication year distributions, but the difference is not huge.

Because the subset of 5,572 works used for regression overlaps heavily with the subset of 10,355 used for longitudinal study, the union of the two produces only 10,830 total volumes.

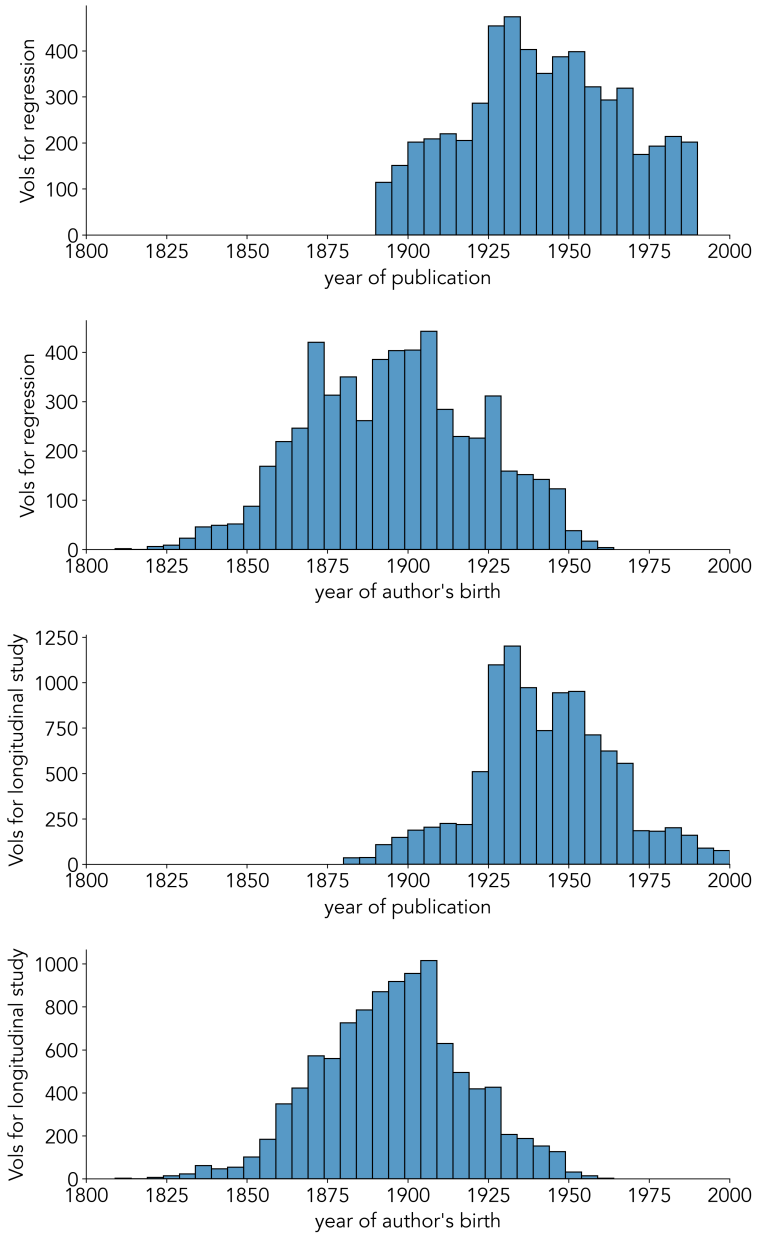

Figure SM 1: The distribution of volumes in two corpora. Each corpus is visualized as a histogram across publication dates, and also authors' dates of birth.

SM 1.3 Text preparation and vocabulary selection

We removed the first 15% and last 5% of pages from all volumes, because front and back matter is rarely composed by the author of the body text. Including it would therefore muffle cohort effects, and bias our model in favor of period effects.

Discarding a fixed fraction of all volumes is certainly an imperfect solution. It doesn't remove all front and back matter. And it comes at the cost of discarding some information about the opening and closing lines of most stories. We feel this is an acceptable cost, since isn't intuitively obvious that the opening or closing lines would be more (or less) strongly shaped by acquired dispositions. But if something like that turned out to be true, this form of data preparation could introduce bias.

In defining a vocabulary for the topic model, we were guided by the advice of Schofield et al. (2017) to be cautious about removing stopwords. We removed only the 26 most common words in the corpus. Schofield advises researchers to make topics interpretable instead by removing a longer list of common words when topic keywords are presented to human readers; we follow this guidance.

We did remove extremely rare words. The metric we used to do this was neither simple term frequency nor document frequency but *author* frequency. Working with a subset of 4,945 authors, we removed words that occurred in fewer than 35 of the writers. The rationale for this approach was to gently discourage the formation of author-specific topics that relied purely on idiosyncratic vocbulary: e.g. "stillsuit" in Frank Herbert. There are more aggressive ways to discourage author-specific topics (Thompson and Mimno, 2018). But that sort of intervention would not be desirable here. De-emphasizing authorial style would have changed the topic model in a way that directly modified a variable of central importance to our experiment: the extent to which an author's works naturally coalesce around a limited number of styles, genres, and themes. That sort of intervention would certainly muffle cohort effects and exaggerate period effects.

Excluding rare vocabulary could, arguably, push in the same direction, slightly reducing the force of cohort effects. But we considered this a worthwhile cost. Readers who are unfamiliar with topic models sometimes dismiss the coherence of topics as "merely a reflection of diction." In defining this model, we wanted to underline that an author's association with specific topics is not simply a consequence of a preference for a few rare words, but emerges organically from theme and genre.

The final vocabulary used in the model (`getEF/modelvocab.tsv`) included 75,451 words.

SM 1.4 Topic categories, and table of topics

Tables 1 through 3 list all the topics in the model we used, along with average delta values and average total r^2 (the amount of variance in topic prominence that can be explained by age, period, or cohort variables). Note that this is prominence in individual books; r^2 would be quite a bit higher if we were only predicting mean topic prominence in a given year.

For much fuller descriptions of topics, and statistics associated with them, see `topic_summary.tsv` in the top level of the code repository in Section SM 3. That file includes a great deal more information about each topic—including a longer list

of key words, a list of the books where the topic is most prominent, the size of the topic, mean word length, and the fraction of documents for which this is the most prominent topic. We consulted all this information before categorizing the topics using `interrater/codingguide.pdf`. We do not expect the labels in Tables SM 1–3 to make intuitive sense given the tiny sample of five keywords you're able to see here.

SM 1.5 Can a topic model register literary form?

At first glance, a "topic model" sounds like a model of content, since the word *topic* usually means a theme or subject. But the implied specificity is misleading. In reality, the algorithms used for modeling just identify selections of language that tend to appear in the same discursive contexts: they don't do anything to separate subjects from kinds of diction associated with styles, formats, or social situations. It is very common, for instance, to see a topic composed of English contractions, because contractions group together in informal contexts. In our model, this is topic 56.

This failure to separate subject matter from social context has posed a problem for some intended applications of topic modeling. For instance, topic models have not functioned well as alternatives to subject headings in library catalogs (Hagedorn et al., 2011). But the promiscuous blending of subject, style, and format can be a positive feature for researchers who are using topic models simply to understand discourse on a macroscopic scale. Topic models do in practice register a wide range of discursive categories: not only subjects and genres, but levels of literary prestige and even formal structures.

In the main article text, we cite some previous research illustrating this principle. To demonstrate how far it reaches, we have run an experiment on the structural distinction between "novel" and "short story," which poses a particularly hard test case. Short stories can be written in a wide range of genres, and on an infinite variety of themes. Our corpus includes collections of stories by Anton Chekhov, Isaac Asimov, and Arthur Conan Doyle, for instance. At first glance, these stories would seem lexically less akin to each to other than they are to Russian realist novels, science fiction novels, and detective novels. It is not immediately obvious why a short work by Asimov would use the same diction as a short work by Anaïs Nin. What they have in common—one might argue—is merely that they're short, and this structural similarity will be erased by a topic model that ignores word order and divides all volumes arbitrarily into chunks of 10,000 words (especially since the chunks will not align with story boundaries).

However, in practice, it is possible to distinguish short story collections from novels using the topic proportions in our model.

The test we ran is documented as `shortstoryexperiment.ipynb` in our code repository. We selected 710 short stories from our corpus by selecting volumes that had the word "other" and the word "stories" in the title. Generally, this matches titles that end with the formula "and other stories." We selected 710 volumes clearly identified as novels by requiring the phrase "a novel" in the title. Since there are many novels in the corpus, we could select examples that precisely matched the

Table SM 1: Statistics for individual topics 0-74

Topic label	category	keywords	delta	total r2
0. War, mostly WWII	event	war german french germans france	0.05	0.048
1. E20c scientists, labs, and professors	genre	professor man quite something thing	0.58	0.017
2. Objects in relation to doors pockets etc	physical description	door put box paper hand	0.64	0.007
3. Late 20c US political thrillers	genre	american people president our security	0.24	0.137
4. Diffuse	uncategorized	man old people men because	0.78	0.028
5. Ambitious European writers	genre	without little eyes same those	0.71	0.018
6. The Western, e20c	genre	trail ranch cattle camp men	0.59	0.051
7. E20c adventure fiction	genre	upon man door moment night	0.25	0.148
8. E20c medicine	technology	doctor dr patient case physician	0.77	0.012
9. Life in organizations, mid-20c	institutions, practices, relationships	office work job man good	0.64	0.028
10. Oral storytelling	uncategorized	went came took told got	0.65	0.014
11. E20c industrial work	technology	men work man mill mine	0.56	0.013
12. Physical sensation, mid-20c	physical description	face eyes voice against felt	0.62	0.11
13. Good society at the end of the 19c	institutions, practices, relationships	young lady most such ladies	0.51	0.194
14. Food and cooking	physical description	eat food kitchen table water	0.85	0.023
15. French language	dialect / language	de madame la le monsieur	0.85	0.012
16. Dialectical representation of African-Am	dialect / language	an de dat ter yo	0.16	0.023
17. First person	uncategorized	myself went saw came told	0.86	0.029
18. Fantasy, late 20c	genre	lord around sword toward dragon	0.13	0.068
19. Mid-20c interiors	physical description	room door around house table	0.43	0.107
20. Subjective thought and feeling	uncategorized	himself thought felt knew how	0.56	0.044
21. Present tense	uncategorized	says does looks comes goes	0.55	0.042
22. E20c rail travel	technology	train station man hotel get	0.22	0.048
23. Horse-drawn modes of transportation	technology	horse horses rode ride riding	0.94	0.018
24. Houses at night, sleep, darkness	physical description	room door bed night house	0.3	0.022
25. First-person plural	uncategorized	our us ourselves each see	0.49	0.011
26. Works older than 1880	dialect / language	upon being having such great	0.41	0.027
27. Ambitious late 20c works in translation	genre	asked just eyes voice suddenly	0.42	0.058
28. First person / dialogue	uncategorized	myself our am us mine	0.6	0.05
29. British rural local colour, e20c	nationalities, regions, or ethnicities	squire old man round good	0.64	0.016
30. China and Japan	nationalities, regions, or ethnicities	chinese japanese china li rice	0.6	0.005
31. Children and families	institutions, practices, relationships	child baby little children woman	0.8	0.011
32. Rural American dialect	dialect / language	got cabin ai reckon wagon	0.79	0.017
33. E20c humorists	genre	says em got n see	0.59	0.046
34. Late 19c romantic historical fiction	genre	am shall may must upon	0.68	0.125
35. Late 20c crime fiction	genre	car got just know maybe	0.12	0.244
36. Late 20c hospitals	technology	hospital nurse doctor ward patient	0.47	0.018
37. Art and artists	institutions, practices, relationships	art picture work artist painting	0.79	0.006
38. Late 19c home life	institutions, practices, relationships	little home good day work	0.9	0.118
39. Landscape description	physical description	rain fire wind light night	0.77	0.041
40. E20c philosophical / social reflection	institutions, practices, relationships	man life its men world	0.41	0.075
41. Sagas and romances	dialect / language	king men knight great castle	0.55	0.017
42. People in groups	uncategorized	men these each people themselves	0.59	0.014
43. Late 20c SF	genre	earth ship space its planet	0.64	0.094
44. 1001 Nights	author-dominated	thou thee thy hath hast	0.28	0.016
45. PG Wodehouse	author-dominated	butler man lt thing yes	0.53	0.008
46. Late-19c sentiment	institutions, practices, relationships	heart poor little face life	0.09	0.161
47. E20c manners	institutions, practices, relationships	mrs mr husband woman wife	0.37	0.024
48. Bars and drinking	institutions, practices, relationships	drink glass bottle bar man	0.79	0.074
49. Biblical language	dialect / language	god lord jesus man shall	0.34	0.009
50. Mid 20c crime fiction	genre	got get just right hell	0.39	0.165
51. Personal interaction	institutions, practices, relationships	re just get even good	0.39	0.078
52. Very general topic	uncategorized	himself though own hand made	0.88	0.019
53. 20c popular stories of marriage	genre	new house room dinner mrs	0.56	0.076
54. E20c young women	institutions, practices, relationships	miss grace girl mr herself	0.37	0.036
55. Late 19c moral obligation	uncategorized	man himself good own such	0.81	0.129
56. Contractions	accident of transcription	don't i'm i've i'll can't	0.46	0.008
57. Giant topic	uncategorized	re know want ca why	0.64	0.043
58. Modernists and experimental writers	uncategorized	am must know think should	0.54	0.019
59. Adverbs and qualifiers	uncategorized	even still just though might	0.38	0.051
60. British city life, mostly London, e20c	nationalities, regions, or ethnicities	london street mr england pounds	0.62	0.004
61. Mid-20c books for children	genre	thought herself went looked old	0.39	0.027
62. Late-19c family life, largely Trollope	author-dominated	sister brother should such must	0.7	0.034
63. Revolutionary and communist movements	event	people party government revolution new	0.27	0.012
64. Informal diction	uncategorized	get re know think going	0.28	0.042
65. Dialogue?	uncategorized	asked looked thought know think	0.67	0.132
66. Money and finance	institutions, practices, relationships	money hundred dollars five thousand	0.85	0.021
67. E20c love and romance	institutions, practices, relationships	girl young love girls man	0.39	0.065
68. Jules Verne and OCR errors	author-dominated	gideon litde french eden swan	0.72	0.001
69. Prison	institutions, practices, relationships	prison mason sheriff cell marshal	0.69	0.006
70. Matter-of-fact journalistic language	uncategorized	upon while being until soon	0.55	0.146
71. Sentimental e20c stories about children	genre	little children christmas papa girls	0.82	0.029
72. Psychological or occult melodrama	uncategorized	herself knew thought might woman	0.38	0.023
73. Religion, mostly Protestant	institutions, practices, relationships	church minister sunday christian reverend	0.43	0.018
74. Diffuse	uncategorized	just say see though looking	0.46	0.014

Table SM 2: Statistics for individual topics, 75-149

Topic label	category	keywords	delta	total r2
75. Passage of time	physical description	day three years night five	0.74	0.023
76. Mid-20c US novels about rural life	uncategorized	got get just around going	0.75	0.045
77. Catholicism in the past	institutions, practices, relationships	bishop dean church monk saint	0.36	0.005
78. Late 19c love stories	genre	am little quite think should	0.88	0.187
79. Cosmopolitan society	institutions, practices, relationships	english paris french american hotel	0.68	0.015
80. Mid-20c nostalgia for an earlier England	genre	quite must good got tea	0.7	0.041
81. Spanish America and the Southwest	nationalities, regions, or ethnicities	don spanish de senor el	0.46	0.003
82. The Middle East	nationalities, regions, or ethnicities	al desert arab sultan allah	0.66	0.003
83. Early 20c popular fiction	genre	know went yes tell room	0.8	0.111
84. America and American history	nationalities, regions, or ethnicities	new washington north york boston	0.46	0.037
85. Native American history and belief	nationalities, regions, or ethnicities	indian indians white chief people	0.84	0.024
86. Warfare, probably centered on WWI	event	captain colonel major lieutenant sergeant	0.56	0.018
87. Late 19c abstract diction	uncategorized	upon might its should himself	0.95	0.196
88. Mid-to-late 20c detective fiction	genre	door know desk office phone	0.38	0.057
89. Crime fiction and gritty urban realism	genre	got get just know yeah	0.61	0.136
90. Verbs of subjectivity	uncategorized	himself thought knew might even	0.61	0.025
91. Gardens and plants	physical description	garden flowers green trees tree	0.72	0.011
92. Rural historical fiction	genre	village old people town man	0.77	0.012
93. British peerage and gentry	institutions, practices, relationships	lady lord earl madam lordship	0.09	0.018
94. German-speaking countries and dialects	nationalities, regions, or ethnicities	herr von baron count german	0.58	0.007
95. Catholic religion	institutions, practices, relationships	priest church st god father	0.42	0.007
96. British dialects	dialect / language	yer im er em ave	0.23	0.004
97. 19c melodrama and sensationalism	genre	upon cried its moment such	0.21	0.174
98. Diffuse / anything	uncategorized	just going people because something	0.2	0.022
99. Feminine protagonist	uncategorized	herself woman girl eyes face	0.58	0.015
100. Westerns or American regionalism	nationalities, regions, or ethnicities	got just around old big	0.37	0.05
101. Restrained, clinical description	physical description	man eyes looked face head	0.58	0.058
102. Tales of US politics	genre	president senator committee vote state	0.52	0.011
103. Family	institutions, practices, relationships	woman old man wife husband	0.72	0.024
104. Sea travel (and perhaps also colonialism	technology	island ship sea captain land	0.86	0.01
105. Stories from South Asia	nationalities, regions, or ethnicities	village house also day even	0.73	0.01
106. Detective fiction	genre	police man inspector murder case	0.25	0.018
107. Hard to label	uncategorized	know himself told thought think	0.29	0.033
108. Love and marriage	institutions, practices, relationships	love life woman heart loved	0.65	0.034
109. Spiritual and philosophic generalization	uncategorized	life world its even still	0.69	0.072
110. War	event	men war enemy line fire	0.4	0.018
111. Writers who are mid-20c women?	uncategorized	think know people just little	0.38	0.016
112. Proper names?	uncategorized	smith brown eve pierce wells	0.86	0.003
113. Late-20c high culture	genre	even though those being old	0.58	0.137
114. E20c poetic diction	uncategorized	upon heart its great love	0.42	0.045
115. Human faces and expressions	physical description	looked turned smiled walked eyes	0.31	0.106
116. American g-droppin' dialects	dialect / language	ai got an em goin	0.72	0.084
117. Mid-century thinky British fiction	genre	thought herself felt seemed looked	0.21	0.025
118. Verbs of speaking asking etc	uncategorized	know mark why asked because	0.27	0.065
119. Violence, esp. mob violence	institutions, practices, relationships	men man god blood face	0.79	0.039
120. Simple diction?	uncategorized	man old himself went came	0.53	0.014
121. Family	institutions, practices, relationships	father son mother old grandfather	0.49	0.006
122. Aeronautics	technology	air plane pilot flying flight	0.46	0.024
123. Genteel comedy	institutions, practices, relationships	mr man young say himself	0.46	0.024
124. Late-20c spirituality and psychology	uncategorized	life because being world human	0.19	0.115
125. Mountains and rough landscapes	physical description	mountain rock valley mountains rocks	0.68	0.025
126. Human bodies and movement	physical description	around head toward hand feet	0.68	0.135
127. Human faces and speech	physical description	eyes face upon voice hand	0.27	0.235
128. Jews and Judaism	nationalities, regions, or ethnicities	jews jew jewish rabbi israel	0.67	0.015
129. Verbs of speech	institutions, practices, relationships	asked replied answered cried exclaimed	0.7	0.153
130. Late-19c realism	genre	mrs oh little know think	0.74	0.135
131. Late-20c satire and cynical comedy	genre	just re really how around	0.29	0.367
132. Dialogue	uncategorized	yes oh know why how	0.65	0.022
133. E20c children's literature and primers	genre	little old great cried good	0.67	0.004
134. Boats and nautical matters	technology	ship captain deck sea boat	0.78	0.007
135. Clothing, fabric, and dress	physical description	white hair black little dress	0.7	0.052
136. Late-19c countryside	physical description	little old its these great	0.83	0.095
137. Schools and teaching	institutions, practices, relationships	school college class teacher students	0.6	0.012
138. Seaside description	physical description	sea water boat beach sand	0.63	0.004
139. Late 20c struggling children	genre	mama just daddy even because	0.81	0.117
140. Difficult to characterize	uncategorized	man came knew know yet	0.75	0.045
141. Late 20c cities	physical description	room street hotel around bus	0.89	0.127
142. Description of late 20c bodies	physical description	its around water eyes light	0.83	0.223
143. Dialogue	uncategorized	am know see must shall	0.63	0.048
144. Fishing and rivers	physical description	river water boat lake fish	0.63	0.008
145. Warfare, probably centered on WWI	event	general men army soldiers war	0.46	0.02
146. Animal stories	genre	dog dogs its fox cat	0.52	0.009
147. Farming	physical description	farm house land barn field	0.77	0.031
148. Inhuman forces	physical description	its upon seemed great yet	0.65	0.032
149. Russia	nationalities, regions, or ethnicities	russian even how russia began	0.79	0.005

Table SM 3: Statistics for individual topics, 150-199

Topic label	category	keywords	delta	total r2
150. Books and publishing	institutions, practices, relationships	book story read books write	0.25	0.008
151. Early 20c metropolis	physical description	street new city york avenue	0.55	0.031
152. Winter weather	physical description	snow ice cold winter wind	0.52	0.019
153. California and the West Coast	nationalities, regions, or ethnicities	san francisco california clay mexican	0.94	0.008
154. Correspondence	institutions, practices, relationships	letter letters read write wrote	0.59	0.042
155. Word segmentation errors	accident of transcription	don ing re con know	0.01	0.148
156. Hunting in the tropics / colonies	institutions, practices, relationships	white bush jungle hut india	0.86	0.005
157. Feelings of spiritual awe	uncategorized	life eyes rose upon seemed	0.21	0.232
158. Money and work	institutions, practices, relationships	money work house good people	0.22	0.021
159. Probably first-person	uncategorized	because myself am say always	0.37	0.013
160. Boys / boyhood	institutions, practices, relationships	boy boys old young little	0.73	0.016
161. The legal thriller: courtroom intrigue	genre	judge court case witness law	0.79	0.009
162. Government	institutions, practices, relationships	general governor government minister chief	0.3	0.008
163. Ireland	nationalities, regions, or ethnicities	irish ireland man dublin sure	0.54	0.005
164. Automobiles	technology	car road driver drive drove	0.12	0.043
165. Banks and finance	institutions, practices, relationships	business money bank victor old	0.58	0.025
166. Tiny topic, proper nouns	uncategorized	grant barker sim gauge shad	0.14	0.005
167. Historical adventure stories	genre	men fort french enemy english	0.23	0.046
168. Uncles, aunts, cousins	institutions, practices, relationships	uncle aunt cousin old house	0.47	0.019
169. Other languages, segmentation errors	accident of transcription	due da na mo ta	0.71	0.001
170. Music and theatre	institutions, practices, relationships	music play stage theatre piano	0.78	0.006
171. Race relations, mostly in the US South	nationalities, regions, or ethnicities	white negro ai got black	0.72	0.019
172. Romance-inflected style	uncategorized	himself even its own against	0.23	0.025
173. Death and evil	uncategorized	dead death man old body	0.47	0.022
174. Stories set in classical antiquity	genre	gods temple city rome caesar	0.5	0.007
175. Historical romance	genre	upon master yet such tis	0.34	0.035
176. Ads at the back	accident of transcription	v cloth crown j c	0.81	0.01
177. France	nationalities, regions, or ethnicities	de madame monsieur m paris	0.48	0.012
178. Marriage	institutions, practices, relationships	married wife marriage husband marry	0.78	0.032
179. Qualifiers	uncategorized	little quite even most once	0.84	0.052
180. Trees and forests	physical description	trees road woods forest tree	0.91	0.009
181. Speech and gestures	physical description	eyes nodded small most appeared	0.71	0.047
182. Deference to nobility	institutions, practices, relationships	sir man may gentleman himself	0.51	0.015
183. Architecture	physical description	house door room wall its	0.53	0.006
184. Folk tales and fairy tales	genre	king princess palace prince son	0.51	0.015
185. Sports stories	genre	game ball play win first	0.68	0.005
186. Italy	nationalities, regions, or ethnicities	florence italian rome italy venice	0.62	0.006
187. Guns	physical description	gun shot man men rifle	0.65	0.021
188. Family relations	institutions, practices, relationships	mother father home family sister	0.33	0.015
189. Informal American families	institutions, practices, relationships	dad ma pa grandma penny	0.83	0.015
190. Hard to say	uncategorized	want know just wanted how	0.55	0.138
191. Human faces and expressions of emotion	physical description	eyes face head hand hands	0.39	0.048
192. Late-20c British social realism?	genre	it round got towards get	0.41	0.018
193. Early 20c boys and dogs?	uncategorized	old little man day half	0.5	0.208
194. Love stories	genre	oh little re just dear	0.32	0.065
195. Measurement	physical description	its these may most same	0.31	0.013
196. Scots dialect	dialect / language	ye an wi man aye	0.67	0.017
197. Historical fiction	genre	king prince queen duke emperor	0.24	0.016
198. Sexuality and erotica	physical description	body bed mouth legs breasts	0.34	0.13
199. Proper names	uncategorized	bill pike april cam rusty	0.45	0.007

dates of our 710 short story collections. (To ascertain whether a topic model can truly discriminate forms, it is important to ensure there is not also a chronological difference between the two sets.) Our text preparation pipeline also excludes paratext by skipping the first 15% and last 5% of pages and ignoring headers that repeat at the tops or bottoms of pages. There should be no explicit references to genre or form in the text we are modeling.

We gathered topic proportions for these 1420 volumes, and repeatedly set 10% of the authors aside as a final test set. We cross-validated a random forest model on the training set to select parameters (number of trees and maximum depth of tree); applying that model to the test set we found that our model of the short story could identify story collections, on average, with 80.1% accuracy. (This figure may be slightly high; to be perfectly accurate we would re-run the 29,341-volume topic model itself each time without the held-out authors, and project it onto the held-out set. But that would take a month to compute, and the difference would be small compared to data variation.)

How is it possible to identify structural categories 80% of the time without word order? For a quick answer we can look at the topics most predictive that a volume is a short story or a novel. It turns out that novels rely heavily on topic 103, "family:

woman old man wife husband young years women children house," as well as topic 75, "passage of time: *day three years night five first four morning last days*." After a moment's reflection, this makes sense. The scope of the novel form obviously allows years to pass, and (less obviously) permits writers to construct sprawling casts of characters connected by a complex web of family relations. In extreme examples, like *Wuthering Heights* and *One Hundred Years of Solitude,* the transformation of a theme as it passes from generation to generation becomes central to the plot.

The topics common in short stories are harder to interpret, but the most predictive of all is topic 157, which we labeled "feelings of spiritual awe: *life eyes rose upon seemed little its face things moment came felt*." These are common abstract words, and the link between them is not completely clear—which is why 157 landed in the "uncategorized" bin. To understand why we nevertheless assigned a label about "spiritual awe," it helps to know that the author most strongly represented in this topic is Algernon Blackwood, a master of the ghost story. But other writers of short stories are also well represented here, and an alternate label could perhaps have been "epiphany": the topic includes vocabulary well suited to describe something that *seemed* or was *felt* in a particular *moment*.

There is in reality no perfect label for a diffuse topic like this. The boundaries of topics are not required to align with the boundaries of familiar concepts. But one of the factors shaping this topic is apparently a rhetoric of intensified, compressed experience that happens to be present in many short stories. If short stories start to outweigh novels in our corpus, this topic will rise and topic 75, "passage of time," will fall.

Our goal here is not to insist on an interpretation of any single topic, but to illustrate generally that literary form is very difficult to separate from content. Short story/novel is ostensibly a purely formal, structural distinction. But a short story doesn't turn into a novel just by describing each event at greater length: the longer form permits writers to describe different aspects of life, and that difference is registered quite clearly by a model of lexical co-occurrence. The same thing is true for differences of literary prestige and perceived quality: topic proportions provide stronger evidence for models discriminating those distinctions than neural document embeddings that try to take account of word order (van Cranenburgh et al., 2019, p. 637).

It is hard to prove a negative: we cannot exhaustively enumerate all possible descriptions of fiction to prove that none of them escape a topic model. The question of whether this mode of representation treats all concepts "equally" is even trickier, because it is difficult to know *a priori* that all boundaries ought to be equally crisp. Perhaps models are very accurate on some categories (like detective fiction) because the pattern in question is governed by strong conventions and are less accurate on others (like the Gothic) because those categories are in reality sprawling and inconsistent.

So we can't prove that topic models introduce no bias at all. But we do have reason to believe that this mode of representation tracks human perception relatively well. One recent experiment shows that the varying accuracy of models trained on lexical evidence correlates with varying human degrees of consensus about the same categories (Calvo Tello, 2021, p. 366). Moreover, topic models seem to

perform well even on test cases we would expect to be particularly difficult—like the abstract boundary between short and long fiction. If lexical models can detect sheer length with 80% accuracy (roughly as well as they can detect the Gothic) there is no reason to assume that they have a blind spot for form. It is of course still possible to envision a scenario where the limitations of topic modeling could be distributed in a way that gave a subtle, systematic advantage to period or to cohort effects. But to construct that scenario we would have to make a series of assumptions that are not supported yet by evidence, or even by intuitive priors.

SM 2 Analytical modeling

SM 2.1 Regression models

We treated the granularity of period and cohort variables as hyperparameters to be optimized, and chose the parameters that minimized squared error on works by out-of-sample authors. In effect this solves a bias-variance tradeoff, adjusting model complexity so each variable contributes as much information as it can without overfitting.

What model specifications did this tend to produce? The granularity of period and cohort factors varied from 4 years to 24 years, creating a 6x6 grid of possible specifications. In figure SM 2, we visualize the number of models in each cell of this grid. The most common specifications are relatively coarse ones: dividing both `birthyear` and `firstpub` variables into 20- or 24-year bins. There are also some models that represent `birthyear` coarsely, but use a 4- or 8-year window for `firstpub`. However, the number of models with a 4-year specification for either variable is low enough to suggest that little would be gained by adding smaller widths.

Figure SM 2 also reports the average r^2 in each cell of the grid. Unsurprisingly, r^2 tends to be higher when at least one of the two variables contained relatively fine-grained information about topic prominence, and was worth dividing into small bins.

In cases where neither variable was strongly predictive, the best-performing strategy might be a cautious, coarse-grained one that minimized overfitting. Thus models with low r^2 tend to concentrate in the lower right. There were many of these, but since we weighted mean δ, across all topics, by r^2 as well as topic size, topics with very low r^2 don't necessarily play a large role in determining the overall average δ. This is why we don't feel much would be gained by adding larger bin widths.

Authors' birth years have a long-tailed distribution, as seen in figure SM 1. A literal application of fixed bin width to those long tails could have produced some 4- or 8-year spans with small single-digit numbers of authors, allowing overfitting. So we allowed bins to widen at the both ends of the timeline.

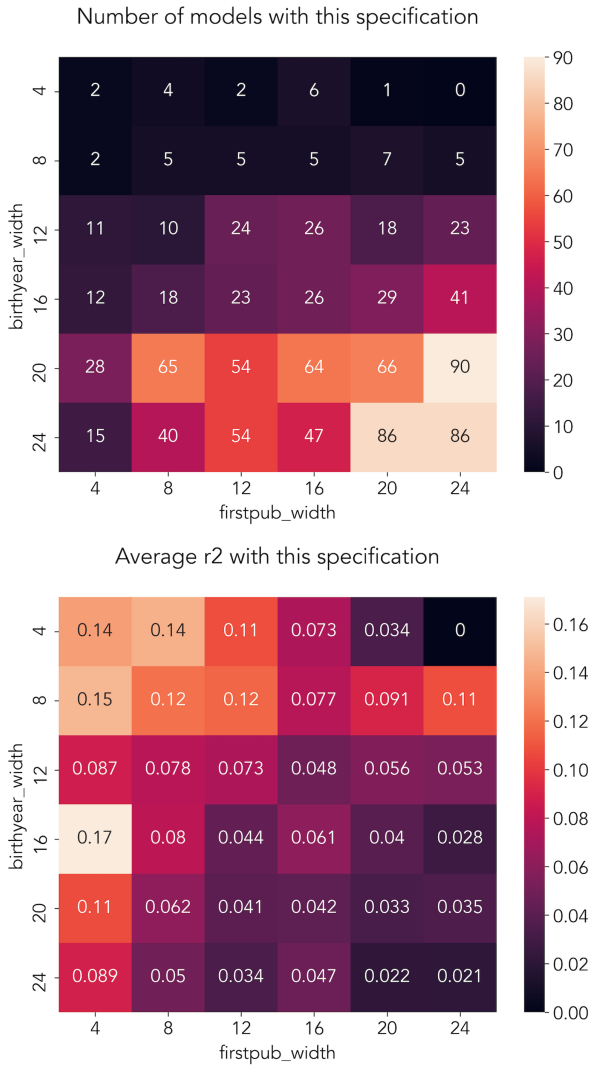

Figure SM 2: The axes on both heatmaps describe the bin width of the factors used to represent authors' birth years (cohorts) or books' first publication years (periods). Since there are 200 topics to model, and we ran the modeling process five times, n = 1000.

SM 2.2 Alternate approaches to regression

We can envision two potential reservations about the strategy described above.

1. Some readers might doubt that discretizing variables does enough to resolve the collinearity of the age-period-cohort triad.

2. Some readers might worry that our grid search hasn't done enough to prevent overfitting.

To address these concerns we've tried several alternate modeling strategies.

First, we can resolve the collinearity question decisively by excluding age from the picture. Age is already playing a small role; it explains only about 6% of the total variance explained by our model. And there are strong *a priori* reasons to doubt that age can do much to explain historical change. A 50-year-old in 1890 may have a body that resembles a 50-year-old in 1980, but their writing styles have little in common.

If we're willing to make the simplifying assumption that age is irrelevant, and remove it, then cohort and period variables are no longer collinear. This addresses concern #1 above, and should satisfy even readers who are highly skeptical of age-period-cohort models. Even Bell and Jones (2013), for instance, who discuss "The impossibility of separating age, period and cohort effects," acknowledge that two of the variables will be separable in cases where we have strong theoretical grounds to believe the third irrelevant.

Running a model without age we find that average delta across all topics is 0.565, very comparable to our reported result of 0.547. Other results in this model are strongly comparable to our main model; our preregistered hypotheses are still confirmed, for instance. See `InterpretAgelessResults.ipynb` in our repository for a fuller discussion.

A second critique, about overfitting, might be prompted by a concern that a 24-year window isn't wide enough to minimize overfitting in some cases. As noted above, we don't see this as a significant problem, because we weighted deltas by r^2, and the total variance explained by time tends to be very low in the lower left corner of figure SM 2. (Large bins were preferred for those topics precisely because time is not very predictive for them.) However, to address concerns, we can return to our primary model (including age), and calculate r^2 only on out-of-sample predictions. We do this by treating r^2 as a machine learning problem: cross-validate the model, and ask how much squared error is added to out-of-sample predictions when we permute the actual values of `birthyear` or of `firstpub`. The results of this alternate strategy are comparable to our primary model. Averaging delta across topics, we get $\delta = 0.493$. Topics are sorted in very much the same order as in our primary model ($r = 0.76$, $p << 0.001$), and our preregistered hypotheses are still confirmed.

We don't believe either of these models is preferable to the one we present in the main article text. Excluding age entirely is a strong assumption, and we don't believe it necessary. Age seems to make a very small difference in literary production, but it could make some difference, and we believe our main strategy has addressed collinearity sufficiently (in the process of addressing other kinds of overfitting). Our second alternate model also seems to us less than optimal, because calculating r^2 through a permutation test is a noisier and more fragile process than ANOVA. We don't think the small risk of overfitting 5,572 volumes with ten 24-year bins actually justifies this added noise.

However, we are reassured to see all three approaches to regression agree that approximately half the variance across a century of literary change is due to

cohort factors. This maximally conservative lower bound on δ would still imply a consequential change to current practice in cultural history.

SM 2.3 Structural equation models

We fit three different structural equation models for each topic: two designed to reflect the durable updating process and one designed to reflect a settled process. The first durable updating model is represented by equations 2-4 in the main text. Broadly speaking, this model assumes that authors have some dispositions or systematic biases in their use of topics, but that changes in style still persist to subsequent works within a set of three works. This model estimates a total of five parameters (ρ, τ, U_i and the variance of U_i and y).

The second durable updating model constrains τ, or the co-variance between individual-level disposition, U_i and the first observed work, y_{i1}, to 0, as well as constraining the variance of U_i to 0. This leaves three parameters to be estimated (ρ, U_i, and the variance of y). This model assumes that authors do not have systematic biases in styles but simply follow random walks with respect to these topics. When the durable change model is preferred, the first, more general model that allows for systematic biases in style, is the preferred model in all but one case (topic 166).

The settled disposition model constrains ρ to 0, meaning that deviations from the baseline, on average, do not translate to a subsequent work. This model has four estimated parameters (τ, U_i, and the variance of U_i and y).

For simplicity, all models constrain ρ to be the same between works 1 and 2 and works 2 and 3. This would be a problematic assumption if these gaps were very different lengths of time. While within authors these gaps of time are highly variable, extending up to 40 years, on average they are very similar. The average gap of time between work 1 and work 2 is 3.14 years, while the average gap of time between work 2 and work 3 is 3.43 years.

Goodness of Fit: We are principally interested in comparing the fit of the settled disposition and durable updating models to each other. Both models are abstractions of distinct causal processes, and as such will likely not explain much of the variation in topic prevalence over time. However, it is important that the preferred models fit the data reasonably well. 50 of the 57 topics (87 percent) preferring the settled dispositions model and 91 of the 128 topics (71 percent) preferring the durable updating model have acceptable fits as measured by the root mean squared error of approximation ($RMSEA < 0.08$). 175 (87.5 percent) of the preferred models have TLI measures above the conventional standard of 0.9.

SM 3 Data availability

The code and data used in the study are archived in Zenodo, and available at `https://doi.org/10.5281/zenodo.5573232`.

The topic model used in the study is archived on Zenodo, and available at `https://doi.org/10.5281/zenodo.5515507`.

References

Bell, A. J. and K. Jones. 2013. "The impossibility of separating age, period and cohort effects." *Social science and medicine* 93:163–165. https://doi.org/10.1016/j.socscimed.2013.04.029.

Calvo Tello, José. 2021. *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. Bielefeld University Press.

Hagedorn, Kat, Michael Kargela, Youn Noh, and David Newman. 2011. "A New Way to Find: Testing the Use of Clustering Topics in Digital Libraries." *DLib Magazine* 17. http://www.dlib.org/dlib/september11/hagedorn/09hagedorn.html.

Schofield, Alexandra, Måns Magnusson, and David Mimno. 2017. "Pulling Out the Stops: Rethinking Stopword Removal for Topic Models." In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 432–436, Valencia, Spain. Association for Computational Linguistics. https://aclanthology.org/E17-2069.

Thompson, Laure and David Mimno. 2018. "Authorless Topic Models: Biasing Models Away from Known Structure." In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3903–3914, Santa Fe, New Mexico, USA. Association for Computational Linguistics. https://aclanthology.org/C18-1329.

Underwood, Ted, Patrick Kimutis, and Jessica Witte. 2020. "NovelTM Datasets for English-Language Fiction, 1700-2009." https://doi.org/10.22148/001c.13147.

van Cranenburgh, A., K. van Dalen-Oskam, and J. van Zundert. 2019. "Vector space explorations of literary language." *Lang Resources and Evaluation* 53:625–650. https://doi.org/10.1007/s10579-018-09442-4.