Journal of Open Humanities Data

The CONLIT Dataset of Contemporary Literature

DATA PAPER

ANDREW PIPER

ubiquity press

ABSTRACT

This dataset includes derived data on a collection of ca. 2,700 books in English published between 2001–2021 and spanning 12 different genres. The data was manually collected to capture popular writing aimed at a range of different readerships across fiction (1,934) and non-fiction (820). Genres include forms of cultural capital (bestsellers, prizewinners, elite book reviews), stylistic affinity (mysteries, science fiction, biography, etc.), and age-level (middle-grade and young adult). The dataset allows researchers to explore the effects of audience, genre, and instrumentality (i.e., fictionality) on the stylistic behavior of authors within the recent past across different classes of professionally published writing.

CORRESPONDING AUTHOR:

Andrew Piper

Department of Languages, Literatures, and Cultures, McGill University, Montréal, CA

andrew.piper@mcgill.ca

KEYWORDS:

literature; fiction; English (language); readership

TO CITE THIS ARTICLE:

Piper, A. (2022). The CONLIT Dataset of Contemporary Literature. *Journal of Open Humanities Data*, 8: 24, pp. 1–7. DOI: https://doi.org/10.5334/johd.88

(1) OVERVIEW

REPOSITORY LOCATION

https://doi.org/10.6084/m9.figshare.21166171.v1

CONTEXT

Access to well-defined collections of contemporary writing is extremely limited today due to intellectual property restrictions, corporate control of data, and the absence of clear consensus surrounding literary categorization. Our dataset is designed to provide researchers with freely accessible derived data of a robust collection of professionally published writing in English produced since 2001, which spans 12 different genre categories. While the term "genre" has been understood in multiple ways within the research community over the years (Cohen, 1986; Underwood, 2016a), we define genre for our purposes as a form of *institutionally framed classification* (Castellano, 2018). According to this definition, genre is what a given institution labels a book using a distinct category of writing.

As we show with the overview of our data (Table 1), our institutional frameworks can include bestseller lists, prize committee shortlists, book review lists, user-generated "choice awards", or corporate forms of categorization. Taken together, they allow research on three different types of institutional framing: cultural capital, stylistic affinity, and reading level. Rather than rely on a single "best" framework, we choose to include multiple forms of selection to allow researchers to explore the effects of different institutional frameworks on stylistic behavior.

In addition to our manually curated selection of books, we also provide researchers with a set of derived features that can be used for further research on the style and content of books (described in Table 2).

Table 1 List of genres, their selection criteria, and the total number of documents per category.

CODE	GENRE	INSTRUMENTALITY	PLATFORM	SELECTION CRITERIA	# DOCS
BIO	Biography	Non-fiction	Goodreads	"Best memoir/biography/autobiography" list	193
BS	Bestseller	Fiction	New York Times	Fiction published since 2001 with the longest aggregate time on the New York Times bestseller list	249
HIST	History	Non-fiction	Amazon	Books listed under "history" under the "bestsellers" tag	205
MEM	Memoir	Non-fiction	Amazon	Books listed under "memoir" under the "bestsellers" tag	229
MID	Middle school	Fiction	Goodreads	Goodreads Choice awards for "Middle Grade" books	166
MIX	Assorted non-fiction	Non-fiction	Amazon	Books listed under assorted non-fiction tags such as "health", "politics", and "business", under the "bestsellers" tag	193
MY	Mystery	Fiction	Amazon	Books listed under "Mystery, Thriller, Suspense" under the "bestsellers" tag	234
NYT	New York Times reviewed	Fiction	New York Times	Fiction reviewed in the New York Times Book Review	419
PW	Prizelists	Fiction	5 Prizelists (US, UK, Canada)	Works shortlisted for the National Book Award (US), PEN/ Faulkner Award (US), Governor General's Award (Canada), Giller Prize (Canada), and the Man Booker Prize (UK)	258
ROM	Romance	Fiction	Amazon	Books listed under "Romance" under the "bestsellers" tag	208
SF	Science-Fiction	Fiction	Amazon	Books listed under "Science Fiction & Fantasy" under the "bestsellers" tag	223
YA	Young Adult	Fiction	Goodreads	Goodreads Choice Awards for Young Adult Fiction	177

(2) METHOD

STEPS

The steps for our dataset construction were the following. Books were manually selected according to the sampling strategies described in Table 1; digitized and manually cleaned; processed using the "large model" of bookNLP (Bamman, 2022); and manually and computationally annotated for features indicated in Table 2.

FEATURE	DESCRIPTION	ANNOTATION TYPE
Category	Fiction or non-fiction	Manual
Genre	Twelve categories	Manual
Publication Date	Date of first publication	Manual
Author Gender	Perceived authorial gender	Manual
POS	Part-of-speech uni- and bigrams	Computational
Supersense	Frequency of 41-word supersenses	Computational
Word Frequencies	Word frequencies for every book/1,000-word passage	Computational
Token Count	Work length measure	Computational
Total Characters	Estimated total number of named characters	Computational
Protagonist Concentration	Percentage of all character mentions by main character	Computational
Avg. Sentence Length	Average length of all sentences per book	Computational
Avg. Word Length	Average length of all words per book	Computational
Tuldava Score	Reading difficulty measure	Computational
Event Count	Estimated number of diegetic events	Computational
Goodreads Avg. Rating	Average user rating on Goodreads	Computational
Goodreads Total Ratings	Total number of ratings on Goodreads as of June 2022	Computational
Average Speed	Measure of narrative pace	Computational
Minimum Speed	Measure of narrative distance	Computational
Volume	Measure of topical heterogeneity	Computational
Circuitousness	Measure of narrative non-linearity	Computational

Table 2 List of 20 features included in our data.

SAMPLING STRATEGY

All books were chosen to represent "popular" writing across 12 different genres of contemporary publishing spanning a 20-year timeframe dating from 2001 through 2021. We define "popular" through multiple criteria that include user-generated awards or lists, elite prize committee lists or book reviews, or bestseller tags on platforms like Amazon or the New York Times. As a further way to validate popularity, we provide two measures drawn from the platform Goodreads.

We define genre through three different kinds of institutional framing: cultural capital (bestsellers, prizewinners, elite book reviews), stylistic affinity (mysteries, science fiction, biography, etc.), and age-level (middle-grade and young adult (YA)). This allows researchers a high degree of flexibility to better understand stylistic behavior of professionally published books targeting different kinds of readerships. We also segment our genres by the "instrumentality" of the information contained ("fiction" or "non-fiction").

While our genre categories are not mutually exclusive (mysteries may appear in Bestsellers and vice versa), no books appear in two separate categories. It is important to note that our larger genre categories (cultural capital, style, age) are not necessarily commensurate with one another and thus researchers should use caution when comparing across these categories. Experimentation with alternative genre labeling systems can be a further affordance of this dataset. Finally, we aimed to select ca. 200 works per category, which we have found is sufficient for training robust text classification algorithms. Due to text availability, list sizes, and cleaning, some categories have more or less than this number. In the case of those books reviewed in the New York Times, we iterated twice on this process. In total, we assemble 2,754 books representing 2,234 unique authors across 12 genres.

To further understand our data, we provide figures of the distribution of publication dates (Figure 1), the average user rating on Goodreads (Figure 2), and the log-transformed number of ratings on Goodreads (Figure 3) to capture book popularity. Finally, while no attention was given to the selection of books based on author gender, our gender distribution across all books

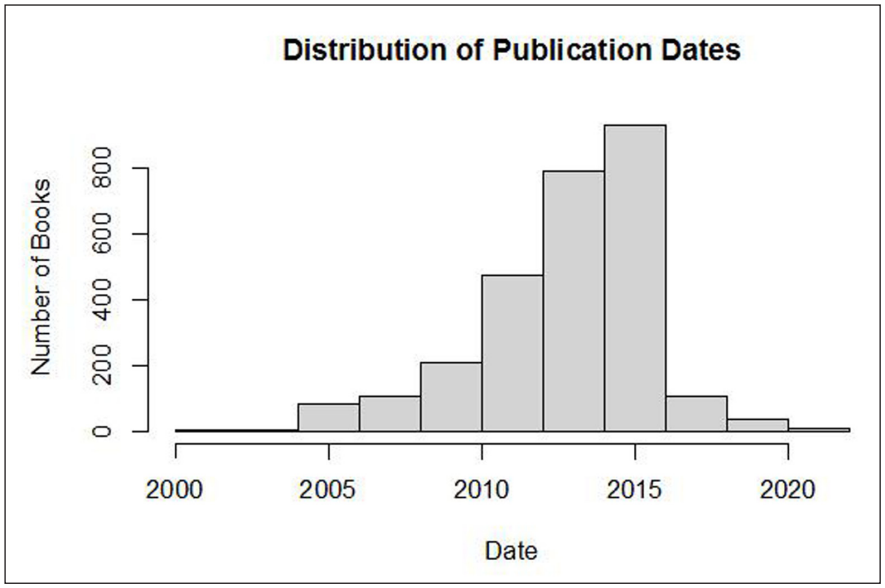

Figure 1 Distribution of publication dates of books in our sample.

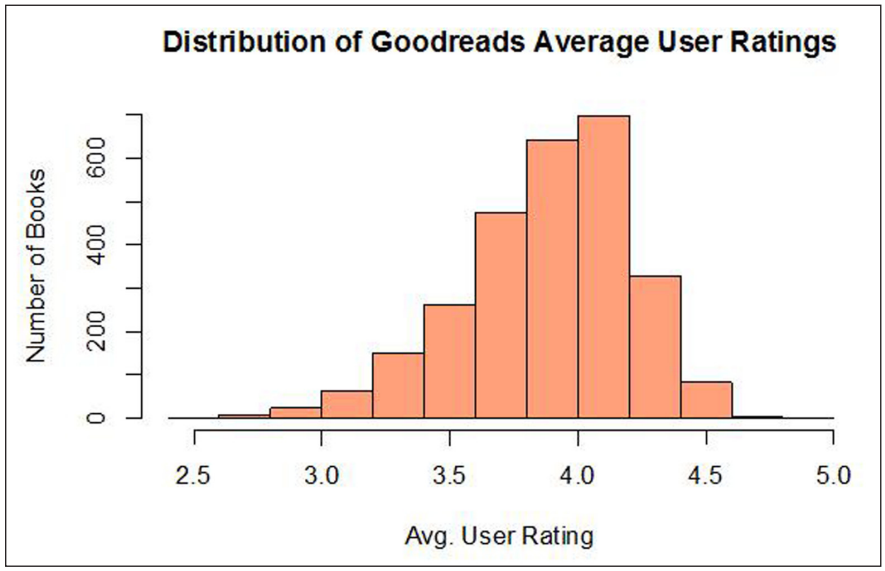

Figure 2 Distribution of the average user rating on Goodreads for books in our sample. Only includes books with > 9 ratings.

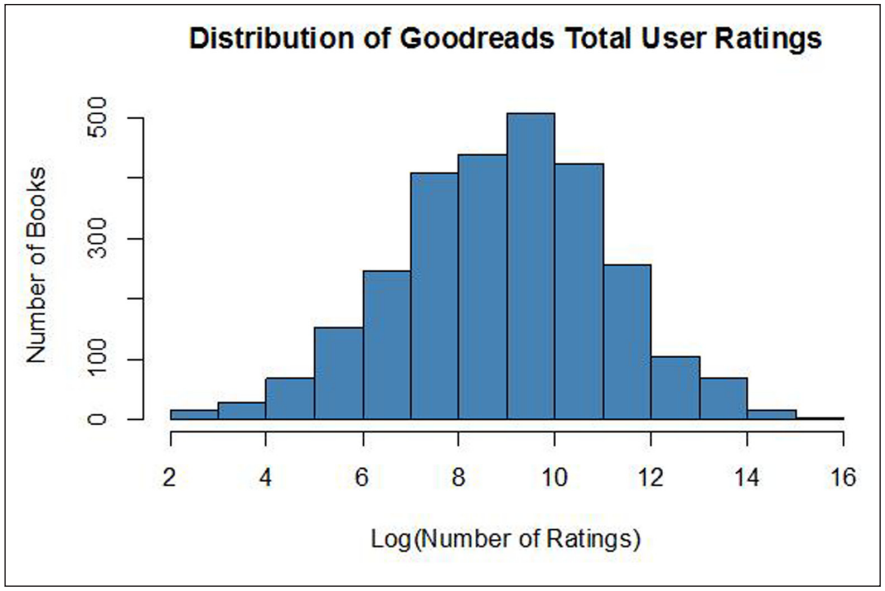

Figure 3 Distribution of the log-transformed number of ratings on Goodreads for books in our sample. Only includes books with > 9 ratings.

is 49.76% women and 49.94% men with only eight books written by self-identified non-binary authors. We note, however, that there are meaningful within-genre differences (Figure 4) as predicted by prior research (Argamon et al., 2003).

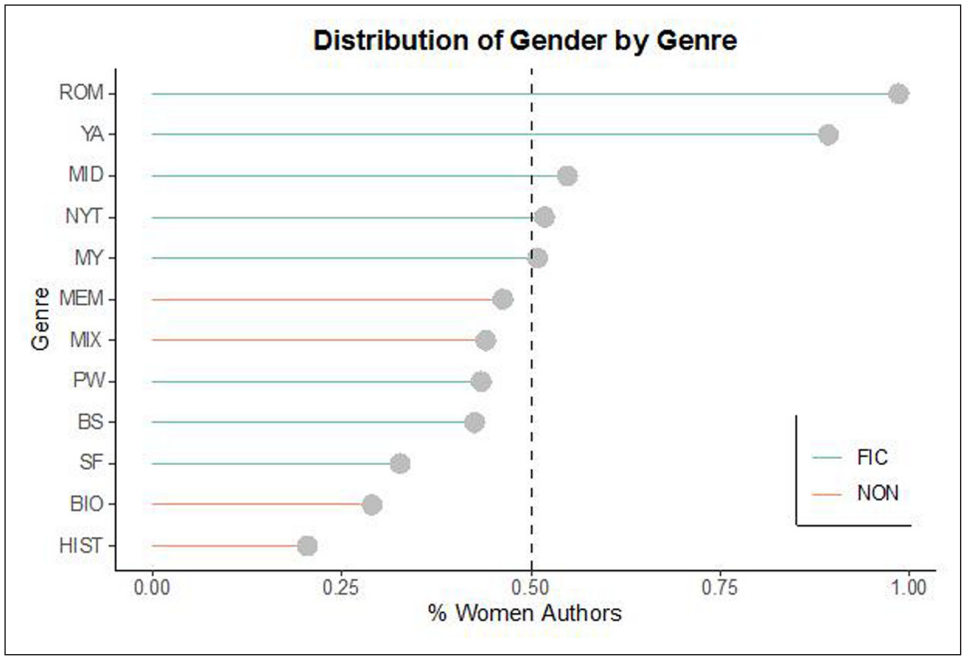

Figure 4 Distribution of author gender by genre.

QUALITY CONTROL

All texts were manually cleaned of front and end matter. Metadata such as publication date, authorial gender, author name and title were all manually entered. The dataset was manually reviewed for the appropriateness of genre labels for every book. Finally, duplicates were removed and any books that were not at least 15,000 tokens in length were also removed. No maximum length was set.

LIMITATIONS

Our data is limited by intellectual property restrictions that do not allow access to full text data. To overcome this limitation, we provide a robust set of derived data that has served in prior research as a reliable foundation for the stylistic understanding of creative writing. Our data is also limited by focusing on a single language. Future work will want to emphasize multilingual data construction to facilitate our understanding of cross-cultural stylistic behavior. Finally, for both manually and computationally derived features, we expect there to be some level of error. For the manual features, we have undertaken two-levels of review. For the computational features, the bookNLP documentation provides estimates on the expected error rates of different predictive models. Nevertheless, it is important for researchers to be aware that our derived features are always estimates. We would flag "Character Count" and "Event Counts" as two features that are worth further research due to the challenging nature of their prediction.

(3) DATASET DESCRIPTION

OBJECT NAME

CONLIT

FORMAT NAMES AND VERSIONS

.CSV

CREATION DATES

Start date: 2015-03-10; End date: 2022-06-22.

DATASET CREATORS

Andrew Piper (McGill University) was responsible for the overall design of the dataset. Eve Kraicer (McGill University) and Joey Love (McGill University) assisted with cleaning and processing the data.

LANGUAGE

English

LICENSE

REPOSITORY NAME

Figshare

PUBLICATION DATE

2022-09-22

(4) REUSE POTENTIAL

Prior work on the computationally driven study of genre has focused on using different selection mechanisms to better understand the role that genre plays in organizing literary communities and reader responses, ranging from studies of historical text data (Sharma et al., 2020; Underwood, 2016b; Wilkens, 2016) to contemporary reader response data (Bourrier et al., 2020; Pianzola et al., 2020; Walsh et al., 2021). Summarizing this work, one could say that research on the content or stylistic aspects of genre has largely focused on historical data while research into contemporary genre formations has largely focused on metadata or non-professionally published writing.

Our dataset is thus designed to give researchers access to stylistic data of contemporary, professionally published writing that spans a range of genre definitions and institutional frameworks. Doing so can help further research into understanding the role genre plays in constraining authorial behavior. It can also facilitate further understanding that the role of differentiation plays in genre classification (Sharma et al., 2022). As genre-theorist Ralph Cohen argued some time ago, "A genre, therefore, is to be understood in relation to other genres, so that its aims and purposes at a particular time are defined by its interrelation with and differentiation from others" (Cohen, 1986, p. 89). Our data will facilitate the empirical exploration of such theories.

By providing Goodreads user response data, our dataset also allows further research into the relationship between style and success (Toubia et al., 2021). The links provided to the Goodreads versions of our books also allow our data to be combined with reader-based response data. An exciting new avenue of literary study aims to better understand the causes and conditions of readers' responses to texts (Mendelman et al., 2021; Pianzola et al., 2020; Walsh et al., 2021) and our data provides the infrastructure to undertake such a research program across a large, diverse set of professionally published contemporary writing.

FUNDING INFORMATION

The creation of this dataset was funded by the Social Sciences and Humanities Research Council of Canada Grant No. 895-2013-1011.

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR AFFILIATION

Andrew Piper orcid.org/0000-0001-9663-5999
Department of Languages, Literatures, and Cultures, McGill University, Montréal, CA

REFERENCES

Argamon, S., Koppel, M., Fine, J., & **Shimoni, A. R.** (2003). Gender, genre, and writing style in formal written texts. *Text & Talk*, *23*(3), 321–346. DOI: https://doi.org/10.1515/text.2003.014

Bamman, D. (2022). *BookNLP*. Retrieved from: https://github.com/booknlp/booknlp (last accessed: June 2022).

Bourrier, K., & **Thelwall, M.** (2020). The social lives of books: Reading Victorian literature on Goodreads. *Journal of Cultural Analytics, 5*(1), 1–34. DOI: https://doi.org/10.22148/001c.12049

Castellano, C. G. (2018). The institution of institutionalism: Difference, universalism and the legacies of Institutional Critique. *Culture, Theory and Critique*, *59*(1), 59–73. DOI: https://doi.org/10.1080/14735784.2017.1410438

Cohen, R. (1986). History and genre. *New Literary History*, *17*(2), 203–218. DOI: https://doi.org/10.2307/468885

Mendelman, L., & **Mukamal, A.** (2021). The generative dissensus of reading the feminist novel, 1995-2020: A computational analysis of interpretive communities. *Journal of Cultural Analytics, 6*(3), 31–73. DOI: https://doi.org/10.22148/001c.30009

Pianzola, F., Rebora, S., & **Lauer, G.** (2020). Wattpad as a resource for literary studies: Quantitative and qualitative examples of the importance of digital social reading and readers' comments in the margins. *PloS one*, *15*(1), e0226708.

Sharma, A., Hu, Y., Wu, P., Shang, W., Singhal, S., & **Underwood, T.** (2020). The Rise and Fall of Genre Differentiation in English-Language Fiction. *CHR 2020: Workshop on Computational Humanities Research*.

Toubia, O., Berger, J., & **Eliashberg, J.** (2021). How quantifying the shape of stories predicts their success. *Proceedings of the National Academy of Sciences*, *118*(26), e2011695118. DOI: https://doi.org/10.1371/journal.pone.0226708

Underwood, T. (2016a). Genre theory and historicism. *Journal of Cultural Analytics*, *2*(2), 1–6. DOI: https://doi.org/10.22148/16.008

Underwood, T. (2016b). The life cycles of genres. *Journal of Cultural Analytics*, *2*(2), 1–25. DOI: https://doi.org/10.22148/16.005

Walsh, M., & **Antoniak, M.** (2021). The Goodreads 'classics': A computational study of readers, Amazon, and crowdsourced amateur criticism. *Journal of Cultural Analytics, 6*(2), 243–287. DOI: https://doi.org/10.22148/001c.22221

Wilkens, M. (2016). Genre, computation, and the varieties of twentieth-century U.S. fiction. *Journal of Cultural Analytics*, *2*(2), 1–24. DOI: https://doi.org/10.22148/16.009

TO CITE THIS ARTICLE:
Piper, A. (2022). The CONLIT Dataset of Contemporary Literature. *Journal of Open Humanities Data*, 8: 24, pp. 1–7. DOI: https://doi.org/10.5334/johd.88

Published: 11 October 2022

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.