Journal of Open
Humanities Data

The CONLIT Dataset of Contemporary Literature

DATA PAPER

ANDREW PIPER

]u[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

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

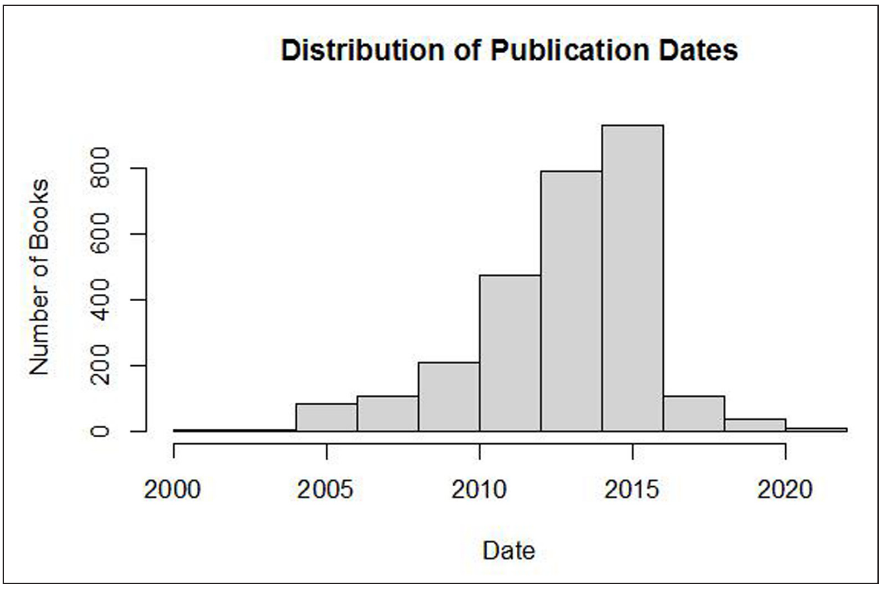

Figure 1 Distribution of publication dates of books in our sample.

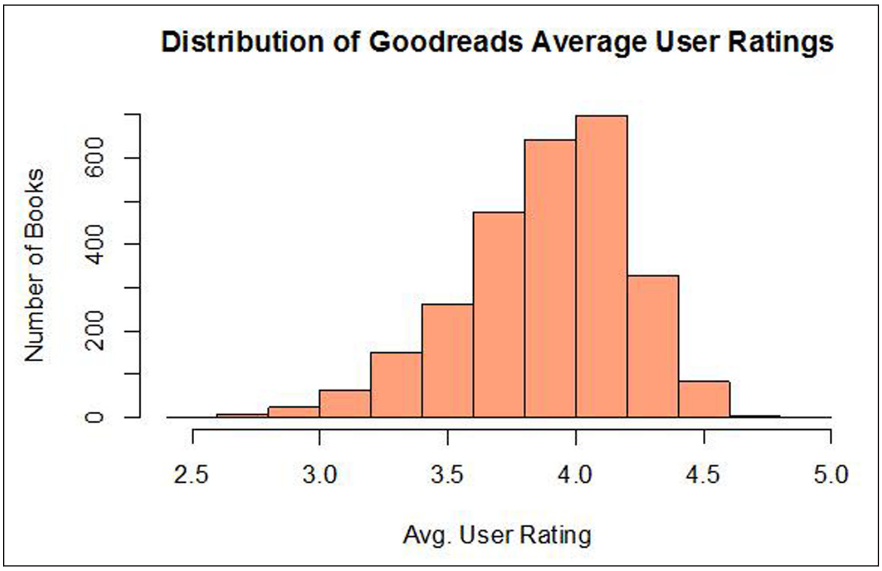

Figure 2 Distribution of the average user rating on Goodreads for books in our sample. Only includes books with > 9 ratings.

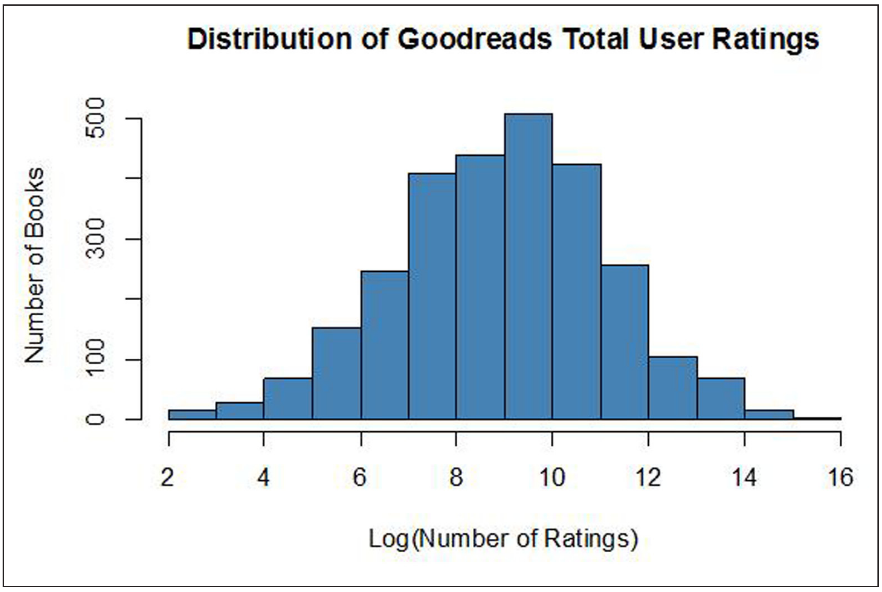

Figure 3 Distribution of the log-transformed number of ratings on Goodreads for books in our sample. Only includes books with > 9 ratings.

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).