



# A Dataset of American Poetry by Poets from Historically Underrepresented Groups in the HathiTrust Digital Library

DATA PAPER

GYURI KANG

KAHYUN CHOI

\*Author affiliations can be found in the back matter of this article

ubiquity press

## ABSTRACT

This dataset provides a collection of American poetry by poets from historically underrepresented groups in the HathiTrust Digital Library. It comprises 9,321 poems from 113 collections by 40 African Americans, 22 Asian Americans, 3 Pacific Islanders, 17 Latin Americans, and 31 Native American poets. We identified and recorded the start and end page numbers for each poem and released the annotations in CSV files. The dataset also reveals imbalances in the representation of poets from historically underrepresented groups within the HathiTrust corpus. We expect this dataset to support large-scale poetry analysis, uncover biases in natural language processing (NLP) models, assess their robustness when applied to culturally diverse poetic language, and promote the development of more inclusive models for diverse American poetry communities.

## CORRESPONDING AUTHOR: Gyuri Kang

Department of Information  
& Library Science, Indiana  
University, Bloomington, USA  
[gyukang@iu.edu](mailto:gyukang@iu.edu)

## KEYWORDS:

American poetry; African  
American poetry; Asian  
American poetry; Pacific  
Islander poetry; Latin  
American poetry; Native  
American poetry

## TO CITE THIS ARTICLE:

Kang, G., & Choi, K. (2026).  
A Dataset of American Poetry  
by Poets from Historically  
Underrepresented Groups in  
the HathiTrust Digital Library.  
*Journal of Open Humanities  
Data*, X(X), pp. 1–6. DOI:  
[https://doi.org/10.5334/  
johd.508](https://doi.org/10.5334/johd.508)

## (1) OVERVIEW

### REPOSITORY LOCATION

CSV files are available via Zenodo (<https://doi.org/10.5281/zenodo.18512641>); Python file and HTRC download instructions are available in the GitHub repository (<https://github.com/krorange/poem-boundary/>).

### CONTEXT

Representing diversity has become increasingly important in digital humanities (DH) research. Beyond English-language literature, many DH studies have examined literary texts in a wide range of languages (Lehmann et al., 2023; Marco et al., 2021; Naaz & Singh, 2022; Saini & Kaur, 2020; Sprugnoli et al., 2023; Timofeeva, 2021; Wessler, 2020). However, even within a single language, literary texts can exhibit substantial variation in style, form, and expression across different ethnic and cultural communities. For example, English is the most widely used language in global book publishing, enabling diverse literary practices across authors and groups.

Despite this diversity, prior DH research on English-language corpora has largely focused on canonical texts by Anglo-American authors. While a small number of studies have begun to examine literary corpora by minority groups in the United States (e.g., Lucy et al., 2025; Parulian et al., 2023; Schug et al., 2025; So, 2020), no existing poetry corpus comprehensively represents multiple marginalized groups in the United States. Creating multicultural corpora can contribute to re-evaluating the English literary canon, which has historically been shaped by racial and gendered ideologies. To address this gap between canonical and non-canonical American literature, we introduce a dataset of American poetry written by poets from historically underrepresented racial and ethnic groups in the United States, including African American (AA), Asian American (APA-AA), Pacific Islander (PA), Latin American (LA), and Native American (NA) poets, sourced from the HathiTrust Digital Library (HathiTrust).

HathiTrust contains approximately 19 million digitized volumes, of which about 8 million are categorized as books. Its holdings have been widely used in DH research spanning fiction, nonfiction, and monographs (Bagga & Piper, 2022; Hamilton & Piper, 2023; Jiang et al., 2021, 2022; Underwood et al., 2020). In contrast, poetry collections within HathiTrust have received comparatively little attention. To assess the coverage of poets from marginalized groups in HathiTrust, we compared the number of poets represented in HathiTrust with those listed on [poets.org](https://poets.org), a poetry website maintained by the nonprofit Academy of American Poets.

Because our analysis focuses on poet-level coverage, we searched for each poet listed on [poets.org](https://poets.org) in HathiTrust. Our results show that HathiTrust includes fewer than half of the poets listed on [poets.org](https://poets.org) across all five groups, with coverage rates ranging from 12.00% to 44.01% (see Table 1). Most groups (AA, APA-AA, and NA) exhibit higher recall rates (above 40.00%), whereas LA shows a lower recall (28.77%) and PA an extremely lower recall (12.00%). These results indicate uneven coverage with distinct representation patterns: while APA-PA and NA are the least represented groups in [poets.org](https://poets.org), LA and APA-PA are the least represented in the HathiTrust.

GROUP	# OF POETS FOUND IN HT	# OF POETS IN POETS.ORG	COVERAGE (%)
AA	136	309	44.01
APA-AA	80	195	41.03
APA-PA	3	25	12.00
LA	42	146	28.77
NA	36	83	43.37

**Table 1** Coverage of poets in HathiTrust, compared with that in [poets.org](https://poets.org).

## (2) METHOD

### STEPS

Based on the list of poets from underrepresented groups compiled by Choi and Kang (2025), we searched for each poet's name in HathiTrust to locate their poetry collections. Because some

poets publish in multiple languages, we restricted our searches to English-language volumes using HathiTrust's language filter to enable consistent comparisons across groups.

For each poet, we collected the volume ID of the most recent poetry collection, as newer volumes tend to provide more reliable metadata and higher OCR quality. Poetry collections were generally easy to identify because most volume titles explicitly include the term *poems* (see Figure 1).

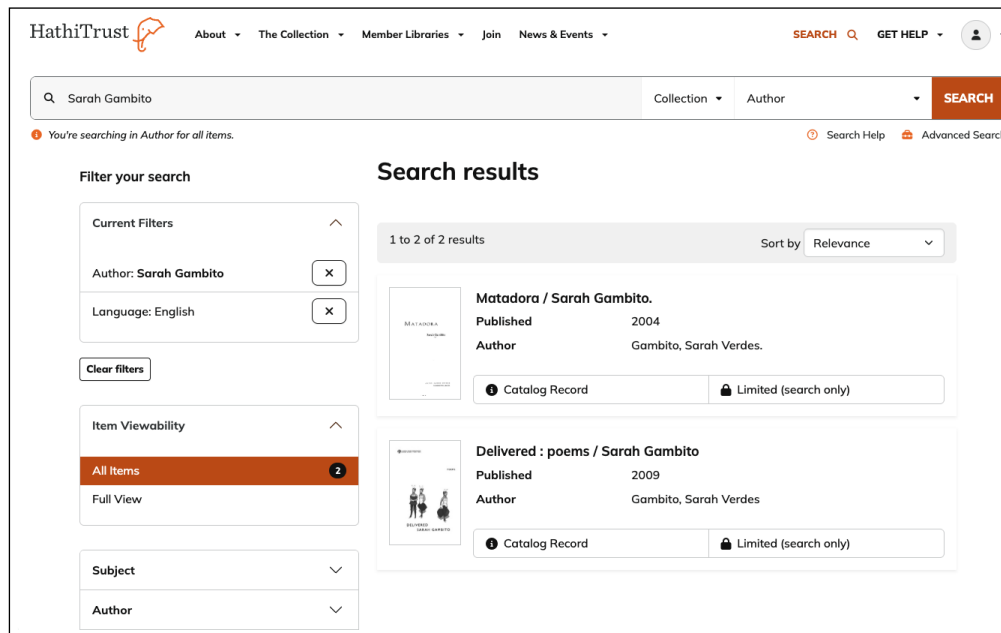


Figure 1 Poetry collection search example in the HathiTrust Digital Library.

In cases where the title did not clearly indicate a poetry collection, we manually inspected the volume by using HathiTrust's "search in the text" feature. Specifically, we searched for terms such as *poems*, *poetry*, and selected poem titles listed on [poets.org](https://poets.org) to verify whether the volume contained poetic content.

Because most volumes in our corpus are under copyright protection, direct access to the full texts is restricted. For research purposes, limited access is available through the HathiTrust Research Center (HTRC) Data Capsules, which provide "secure computing environments for performing researcher-driven text analysis on the HathiTrust corpus" (HathiTrust Research Center, n.d.). Using a Data Capsule, we downloaded the selected volumes and extracted them as individual page-level text files.

For each volume, we manually identified poems that appeared on a single page. Pages containing multiple poems were excluded to ensure consistent sectionalization across the dataset. In addition to the page numbers of digitized copies in the Data Capsule, we include page numbers of OCR-scanned print books when available for users who may consult print copies.

## SAMPLING STRATEGY

Our selection of poets follows the groupings proposed by Choi and Kang (2025), who annotated and published race and ethnicity metadata for poets in an American poetry collection curated by [poets.org](https://poets.org). Based on descriptive tags used by [poets.org](https://poets.org), including designations associated with cultural and heritage observances such as Asian/Pacific American Heritage Month, Black History Month, and Native American Heritage Month, they identified the five groups. Multiracial poets are counted in multiple groups to reflect their affiliation with more than one ethnic community.

Due to variation in availability across groups within HathiTrust, the number of accessible volumes and poems differs by group. Within a defined time period, we randomly selected poetry volumes and manually annotated poem boundaries to identify poems within each volume. To improve representation among groups with smaller holdings, we prioritized annotation for Pacific Islander and Native American poets. Following boundary annotations, we extracted a diverse set of poems across the selected volumes.

To better understand the dataset, we provide summary statistics on the number of available volumes, boundary-annotated volumes, and identified poems for each group (see [Table 2](#)).

GROUP	# OF VOLUMES	# OF BOUNDARY-ANNOTATED VOLUMES	# OF IDENTIFIED POEMS
AA	136	40	3,380
APA-AA	80	22	1,660
APA-PA	3	3	298
LA	42	17	1,563
NA	36	31	2,420

**Table 2** Number of volumes and poems per group.

African American poets constitute the largest group in the dataset: we annotated poem boundaries for 40 of the 141 available volumes and identified 3,380 poems. Pacific Islander poets form the smallest group, represented by three volumes in HathiTrust, resulting in a total of 298 poems.

## QUALITY CONTROL

The group assignments for underrepresented poets were manually verified using descriptive tags from [poets.org](#), following the procedure outlined in Choi and Kang (2025), and supplemented with targeted web searches. During this verification process, we consulted representative websites, including poets' official websites, Wikipedia, and the Poetry Foundation, to confirm that each poet was categorized into the appropriate racial and ethnic groups.

## (3) DATASET DESCRIPTION

### REPOSITORY NAME

A Dataset of American Poetry by Poets from Historically Underrepresented Groups in the HathiTrust Digital Library.

### OBJECT NAME

htrc\_poetry\_sections/aa\_poets  
htrc\_poetry\_sections/apa-aa\_poets  
htrc\_poetry\_sections/apa-pa\_poets  
htrc\_poetry\_sections/lxa\_poets  
htrc\_poetry\_sections/na\_poets  
poem-boundary.py

### FORMAT NAMES AND VERSIONS

CSV files and a Python file

### CREATION DATES

Start date: 2024-06-01 End date: 2025-11-15

### DATASET CREATORS

Gyuri Kang (Indiana University Bloomington); Kahyun Choi (University of Illinois Urbana-Champaign).

### LANGUAGE

English

### LICENSE

CC BY 4.0

## (4) REUSE POTENTIAL

This dataset is designed to support computational analyses of American poetry and to address gaps in computational literary research. Prior DH efforts have sought to represent “historically under-resourced and marginalized textual communities” by constructing datasets focused on African American literature, Native American texts, Black fantastic writing, Latin American fiction, and African American health documents (HathiTrust Research Center, n.d.). Such datasets have been used to evaluate the effectiveness of text mining tools, such as Named Entity Recognition, on non-canonical literary texts (Parulian et al., 2023).

Beyond tool evaluation, the dataset enables the investigation of distinctive linguistic, stylistic, and thematic patterns across racial and ethnic groups of American poets. Future research may employ the dataset for large-scale comparative analyses across groups or for focused studies of specific communities using a range of computational methods. Existing computational poetry research has examined structural aspects of poetry, including rhythm, meter, and stanza (Marco et al., 2021; Naaz & Singh, 2022), syntactic and semantic features (Shang & Underwood, 2024; Timofeeva, 2021), and sentiment and emotion in poetic language (Saini & Kaur, 2020; Sprugnoli et al., 2023).

Several limitations should be considered when reusing this dataset. Firstly, some racial and ethnic groups of American poets are not included. Group selection was informed by comparisons between poets.org data and the 2020 U.S. Census (Choi & Kang, 2025), and therefore, the dataset does not represent all racial and ethnic communities in the United States. Incorporating additional groups, such as Alaskan American or Arab American poets, would improve the dataset’s coverage and representativeness. Secondly, variation in corpus size across groups may introduce bias in comparative analyses. We recommend balancing the number of poems per group or applying appropriate normalization techniques when conducting cross-group comparisons.

## FUNDING STATEMENT

This work was supported by the Institute of Museum and Library Services (Grant No. RE-252382-OLS-22) under the Laura Bush 21st Century Librarian Program.

## COMPETING INTERESTS

The authors have no competing interests to declare.

## AUTHOR CONTRIBUTIONS

Gyuri Kang: data curation, software, formal analysis, investigation, visualization, methodology, writing – original draft, review & editing.

Kahyun Choi: conceptualization, data curation, methodology, investigation, writing – original draft (partial), review & editing, supervision, project administration, funding acquisition.

## AUTHOR AFFILIATIONS

Gyuri Kang  [orcid.org/0000-0002-6132-758X](https://orcid.org/0000-0002-6132-758X)

Department of Information & Library Science, Indiana University, Bloomington, USA

Kahyun Choi  [orcid.org/0000-0003-4854-7104](https://orcid.org/0000-0003-4854-7104)

School of Information Sciences, University of Illinois, Urbana-Champaign, USA

## REFERENCES

Bagga, S., & Piper, A. (2022). HATHI 1M: Introducing a Million Page Historical Prose Dataset in English from the Hathi Trust. *Journal of Open Humanities Data*, 8, 7. <https://doi.org/10.5334/johd.71>

- Choi, K., & Kang, G. (2025). An analysis of poet demographic and thematic diversity in a poetry collection for inclusive AI. *Information Research an International Electronic Journal*, 30(iConf), 610–617. <https://doi.org/10.47989/ir30iConf47263>
- Hamilton, S., & Piper, A. (2023). MultiHATHI: A Complete Collection of Multilingual Prose Fiction in the HathiTrust Digital Library. *Journal of Open Humanities Data*, 9, 3. <https://doi.org/10.5334/johd.95>
- HathiTrust Research Center. (n.d.). *Data capsules*. <https://analytics.hathitrust.org/staticcapsules>
- Jiang, M., Dubniecek, R. C., Worthey, G., Underwood, T., & Downie, J. S. (2022). A prototype gutenberghathiTrust sentence-level parallel corpus for OCR error analysis: Pilot investigations. *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, 1–5. <https://doi.org/10.1145/3529372.3533298>
- Jiang, M., Hu, Y., Worthey, G., Capitanu, B., Kudeki, D., & Downie, J. S. (2021). *The Gutenberg-HathiTrust Parallel Corpus: A Real-World Dataset for Noise Investigation in Uncorrected OCR Texts*. iConference 2021. <http://hdl.handle.net/2142/109695>
- Lehmann, M., Heumann, A., Kuijpers, M. M., Lauer, G., & Lüdtke, J. (2023). The ChildPoeDE Corpus: 1082 German Children’s Poems for Computational and Experimental Studies on Poetry Reception. *Journal of Open Humanities Data*, 9, 6. <https://doi.org/10.5334/johd.102>
- Lucy, L., Griffiths, C., Ying, C., Kim-Ebio, J., Baur, S., Levine, S., Eberhardt, J., Bamman, D., & Demszky, D. (2025). Racial and Ethnic Representation in Literature Taught in US High Schools. *Journal of Cultural Analytics* 10(1). <https://doi.org/10.22148/001c.131682>
- Marco, G., De La Rosa, J., Gonzalo, J., Ros, S., & Gonzalez-Blanco, E. (2021). Automated Metric Analysis of Spanish Poetry: Two Complementary Approaches. *IEEE Access*, 9, 51734–51746. <https://doi.org/10.1109/ACCESS.2021.3069635>
- Naaz, K., & Singh, N. K. (2022). Design and Development of Computational Tools for Analyzing Elements of Hindi Poetry. *IEEE Access*, 10, 97733–97747. <https://doi.org/10.1109/ACCESS.2022.3204388>
- Parulian, N. N., Dubniecek, R., Evans, D. J., Hu, Y., Layne-Worthey, G., Downie, J. S., Heaton, R., Lu, K., Orr, R. I., Magni, I., & Walsh, J. A. (2023). Tuning Out the Noise: Benchmarking Entity Extraction for Digitized Native American Literature. *Proceedings of the Association for Information Science and Technology*, 60(1), 681–685. <https://doi.org/10.1002/pra2.839>
- Saini, J. R., & Kaur, J. (2020). Kāvi: An Annotated Corpus of Punjabi Poetry with Emotion Detection Based on ‘Navrasa.’ *Procedia Computer Science*, 167, 1220–1229. <https://doi.org/10.1016/j.procs.2020.03.436>
- Schug, J., Gosin, M., & Alt, N. P. (2025). A historical psychology approach to gendered racial stereotypes: An examination of a multi-million book sample of 20th century texts. *Current Research in Ecological and Social Psychology*, 9. <https://doi.org/10.1016/j.cresp.2025.100248>
- Shang, W., & Underwood, T. (2024). Disentangling semantic and prosodic features of English poetry. *Digital Scholarship in the Humanities*, fqae008. <https://doi.org/10.1093/lilc/fqae008>
- So, R. J. (2020). *Redlining culture: A data history of racial inequality and postwar fiction*. Columbia University Press. <https://doi.org/10.7312/so--19772>
- Sprugnoli, R., Mambrini, F., Passarotti, M., & Moretti, G. (2023). The Sentiment of Latin Poetry. Annotation and Automatic Analysis of the Odes of Horace. *Italian Journal of Computational Linguistics*, 9(1). <https://doi.org/10.4000/ijcol.1125>
- Timofeeva, M. (2021). Comparative Analysis of Reasoning in Russian Classic Poetry. *Applied Sciences*, 11(18), 8665. <https://doi.org/10.3390/app11188665>
- Underwood, T., Kimutis, P., & Witte, J. (2020). NovelTM Datasets for English-Language Fiction, 1700–2009. *Journal of Cultural Analytics*, 5(2). <https://doi.org/10.22148/001c.13147>
- Wessler, H. (2020). From marginalisation to rediscovery of identity: Dalit and Adivasi voices in Hindi literature. *Studia Neophilologica*, 92(2), 159–174. <https://doi.org/10.1080/00393274.2020.1751703>

**TO CITE THIS ARTICLE:**

Kang, G., & Choi, K. (2026). A Dataset of American Poetry by Poets from Historically Underrepresented Groups in the HathiTrust Digital Library. *Journal of Open Humanities Data*, X(X), pp. 1–6. DOI: <https://doi.org/10.5334/johd.508>

**Submitted:** 07 January 2026

**Accepted:** 14 February 2026

**Published:** XX Month 202X

**COPYRIGHT:**

© 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

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

## Typesetting queries

1. Please ensure that all required competing interests have been declared, including whether any of the listed authors are currently members of the journal's editorial team or board, or have held such a position within the last three years, or confirm that all authors of this article have no competing interests. Further information is available here: <https://openhumanitiesdata.metajnl.com/about/competinginterests/>.
2. In order to increase accessibility of the article, we need to add alt-text to the images. Please provide us with a very short description of the images to be added as alt-text (ideally no longer than 10 words). See this guide for some useful tips on how to write a good alt-text: <https://moz.com/learn/seo/alt-text>