Journal of Open Humanities Data

MultiHATHI: A Complete Collection of Multilingual Prose Fiction in the HathiTrust Digital Library

DATA PAPER

ubiquity press

SIL HAMILTON

ANDREW PIPER

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

ABSTRACT

This dataset provides detailed metadata on ca. 10.2 million works of fiction and non-fiction written after 1799 in 521 different languages available in the HathiTrust Digital Library. The dataset bolsters the May 2022 Hathifile by supplying missing predicted fiction tags with a bespoke BERT-based multilingual classifier. Our classifier completes the catalogue with an additional 400,000 non-English volumes predicted to be works of fiction, capturing 95% of all works presently provided by HathiTrust. We provide each work with metadata including the work's genre at the level of fiction or non-fiction, length in pages, original language, and the year the work was published. With a total page count of ca. 1.4 billion pages, our dataset provides researchers with a substantial source of non-English modern literature. We also present insight into how multilingual classifiers can be trained with monolingual data, itself a discovery with implications for the study of lower resource languages. We hope our provisions will accelerate empirical research into non-English prose and literature.

CORRESPONDING AUTHOR:

Sil Hamilton

Languages, Literatures, and Cultures, McGill University, Montreal, Canada

sil.hamilton@mcgill.ca

KEYWORDS:

fiction; multilingual fiction; non-English prose; world literature

TO CITE THIS ARTICLE:

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, pp. 1–7. DOI: https://doi.org/10.5334/johd.95

(1) OVERVIEW

CONTEXT

Digital Humanities researchers interested in studying fiction face hurdles in acquiring sufficient quantities of non-English literature on which to conduct their experiments. While digital heritage collections like HathiTrust have made progress in providing researchers with meaningful access to non-English literature, works written in languages other than English continue to suffer from disproportionately more metadata and accessibility issues than works written in English (Mahony, 2018; Bagga & Piper, 2022; Fenlon et al., 2014).

The prevalence of low quality metadata is apparent when surveying metadata statistics on the HathiTrust Digital Library, where non-English works are more likely to be missing (non-) fiction tags relative to works written in English (Figure 1). A discrepancy of this magnitude has ramifications for studies seeking to incorporate the most general distribution of written materials possible: if 10% (or more) of available texts in a given language are effectively unavailable because of metadata issues, the potentialities afforded to the humanities by big data are hampered (Van Eijnatten, 2013). This danger is exacerbated given non-English texts make up 47% of HathiTrust's catalogue.

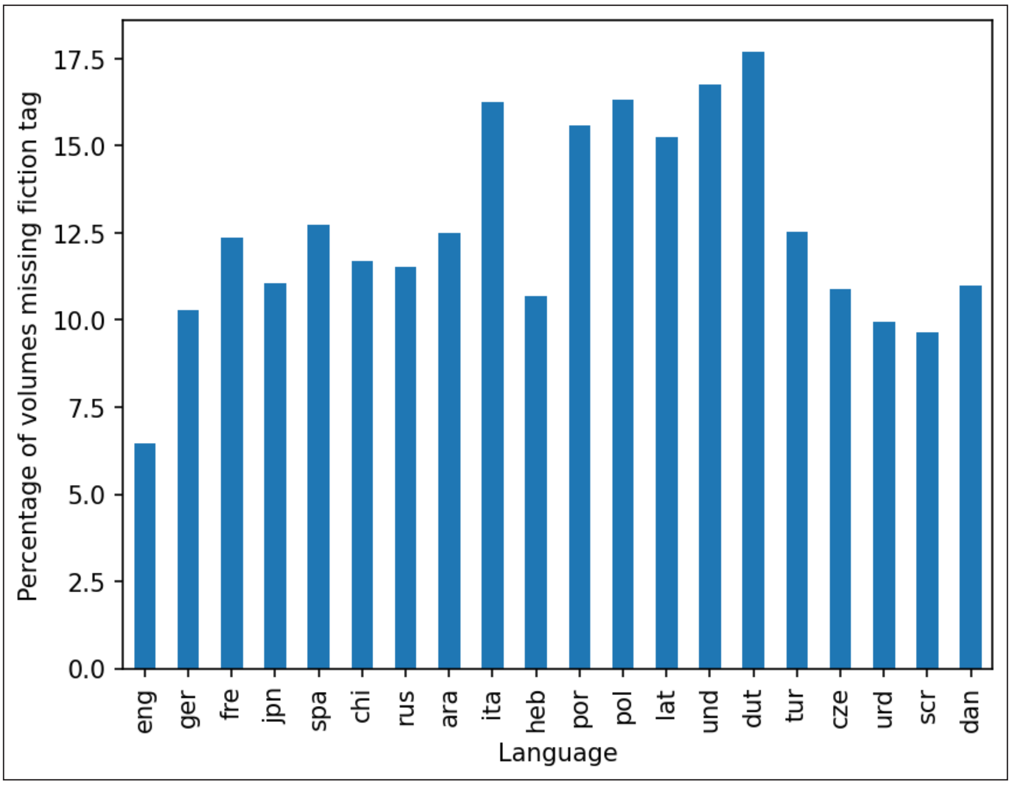

Figure 1 Percentage of books missing a fiction tag for the 20 most frequent languages.

We seek to provide this missing metadata with a bespoke fictionality classifier trained on monolingual data. Fictionality for our purposes is an institutionally-defined classification indicating whether a work is intended to be fictional or not, a classification rendered through how a work is written (Ryan, 1980). The resulting dataset provides quantitative researchers with a complete list of volumes predicted to be fictional or non-fictional available in the HathiTrust Digital Library. We increase the total number of tagged works in HathiTrust up from ca. 9.7 million to ca. 10.2 million works, totaling an increase of ca. 400,000 works. Our final dataset captures 95% of all works provided by HathiTrust. To ease researcher access into this collection, we provide additional metadata which allows for easy subsetting of the global list according to individual researcher preferences (Table 1).

ATTRIBUTE	DESCRIPTION
HTID	The HathiTrust ID by which the work is accessible.
Access Restrictions	Whether the work is made public by HathiTrust.
HathiTrust Bibliography Key	The respective bibliography key for the work. For retrieving MARC records.
Title	The title of the volume in question.
Year Published	The year in which the work was published.
Language	The language in which the work was published.
Author	The author of the work in question.
Fictionality	Whether the work is intended to be fictional (1) or not (0).
Length	The length of the work.

Table 1 List of attributes included in our dataset.

(2) METHOD

STEPS

We began constructing our dataset by obtaining an extant list of works currently provided by HathiTrust. This Hathifile lists all volumes made available on the HathiTrust Digital Library platform. We cleaned the dataset and proceeded to download bibliographic data for every single entry. We subsetted the dataset for works missing fiction tags and pass these on to our classification process running on the HTRC Data Capsule. We then reintegrated the predicted values. A detailed description of this classification process follows.

CLASSIFICATION PROCESS

We used a Transformer-based classification process to predict missing fiction tags. Given our goal was to increase the number of classified non-English works in the Hathifile, we required a multilingual bidirectional encoder. We selected XLM-RoBERTa (base) for this task given it was pretrained on substantially more literary data than competing models and thus remains the state of the art in cross-lingual language models (Conneau et al., 2019).

We equipped XLM-RoBERTa with an additional classification layer and trained this layer for five epochs on 144,000 examples of 512-word spans of English fiction and non-fiction drawn from the CONLIT dataset (Piper, 2022). We assessed model performance on novels written in ten different languages as contained in the European Literary Text Collection (Odebrecht, C., Burnard, L., & Schöch, C., 2021) and private non-fiction corpora consisting of textbooks and biographies. Further tests were conducted on private non-fiction German, Japanese, and French corpora. We found that our model performs well (minimum 80% F1-score) in all tests despite having only been trained with English samples (see Quality Control for further details). We quantized the model to improve classification speed in the data capsule (ONNX Runtime Developers, 2021).

Moving the model into the HTRC Data Capsule, we proceeded with downloading ten random pages from each volume to be classified. Given ten classifications per volume, we used majority vote to assign the volume an overall fictionality tag. We exported the predicted ca. 400,000 tags and re-integrated the data into the overall Hathifile. As can be seen in Figure 2, our process significantly improves the metadata coverage of non-English languages in the Hathi catalogue.

When examining the temporal distribution of non-English texts in the HathiTrust Digital Library (Figure 3), we find our classification process provides the researcher with substantially more texts published during the post-war period. While this data is copyrighted and thus access is limited by HathiTrust, researchers interested in studying non-English contemporary fiction will find HathiTrust has more to offer than one may previously have believed due to incomplete metadata.

QUALITY CONTROL

We sampled and tested classified works to ensure accuracy. To assess the quality of our classifications, we provided native speakers of ten different languages with titles and pages from twenty random works of fiction and non-fiction written in their respective languages. This process indicates our RoBERTa-based classifier achieves a harmonized F1-score of 88.9%. We provide per-language scores in Table 2. We then cleaned and de-duplicated the dataset to ensure label consistency.

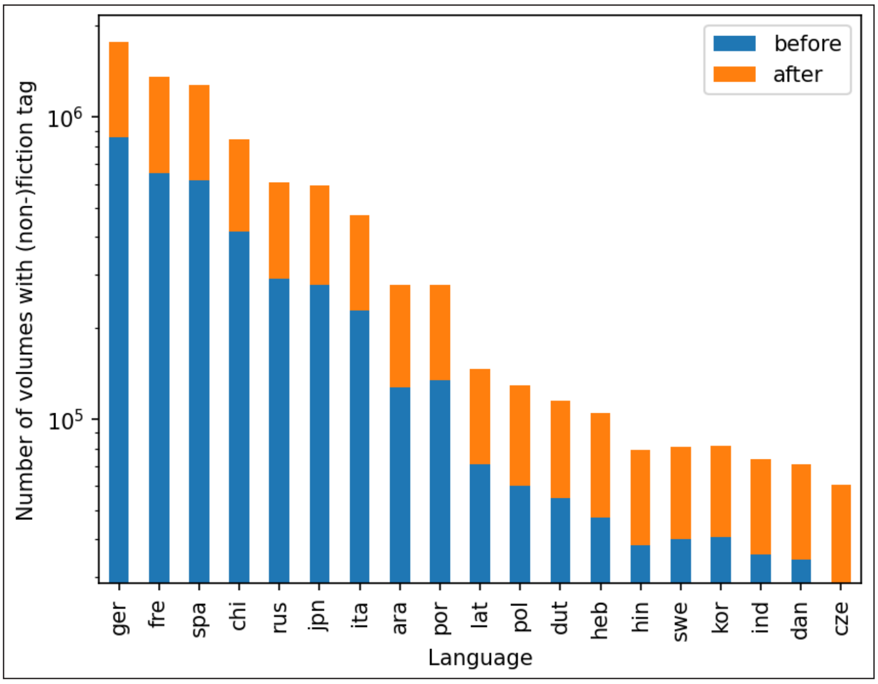

Figure 2 Number of books tagged as fiction for the 18 most frequent languages, before and after classification.

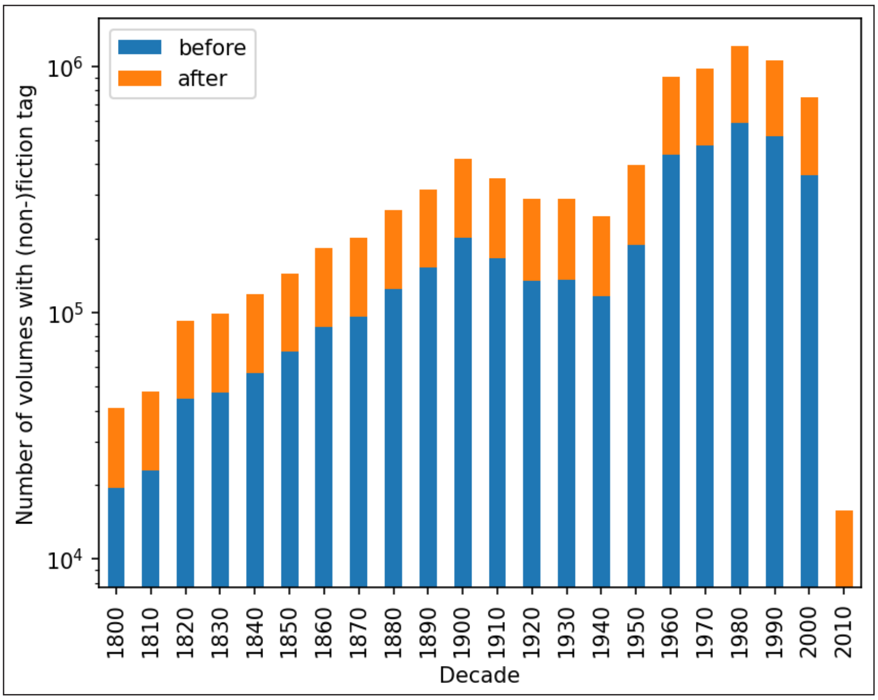

Figure 3 Relative number of non-English books by decade before and after classification.

Table 2 List of evaluated languages and their respective precision, recall, and F1 scores.

LANGUAGE	PRECISION	RECALL	F1
German	80%	88%	84%
Italian	100%	90%	95%
Japanese	100%	90%	95%
Russian	90%	90%	90%
Dutch	80%	100%	88%
Hebrew	80%	100%	88%
Danish	100%	76%	87%
Chinese	100%	83%	91%
Arabic	50%	100%	66%
Polish	90%	100%	94%

LIMITATIONS

Our dataset is limited by access restrictions. While many of the works are accessible via public APIs provided by HathiTrust, the majority of written fiction published after 1923 is only accessible through data capsules through the HathiTrust due to intellectual property restrictions in the United States. We account for this limitation by indicating in the metadata whether a given volume is accessible to the public. A further limitation concerns the imperfect nature of our classification algorithm. While the overall harmonized score for our classifier is high, we note we only tested the ten most frequently used languages in the dataset. Researchers interested in data that is 100% accurate can use our data as a means of reducing the burden of manually curating data sets. At the same time, prior research has shown that imperfectly classified data can be used for large-scale inferences of cultural behavior (Bagga & Piper, 2022; Underwood, 2014).

(3) DATASET DESCRIPTION

OBJECT NAME

MultiHATHI

FORMAT NAMES AND VERSIONS

.CSV

CREATION DATES

Start date: 2022–05–01 End date: 2022–10–30

DATASET CREATORS

Sil Hamilton and Andrew Piper

LANGUAGE

English, German, French, Spanish, Italian, Russian, Japanese, Chinese, more.

LICENSE

Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)

REPOSITORY NAME

FigShare

PUBLICATION DATE

01/12/2022

(4) REUSE POTENTIAL

Prior work in the Digital Humanities has highlighted the importance of multilingual corpora for cultural study (Mahony, 2018; Spence & Brandao, 2021; Gil & Ortega, 2016). In treating this absence, we present researchers with a dependable list of volumes with predicted fictionality tags representing over 500 languages. In doing so, we significantly increase the number of readily available (non-)fictional texts currently provided by the HathiTrust Digital Library. This dataset can be used to further our understanding of categories like 'fictionality' (Piper, 2022), 'narrativity' (Bagga & Piper, 2022), 'genre' (Underwood, 2014), 'place' (Evans, 2018) across numerous cultural contexts beyond English. At the same time, it can be useful for the NLP community in search of multilingual data sets for genre-specific or linguistic prediction tasks (Ogueji, 2021).

We also wish to underscore the viability of training multilingual Transformer-based classifiers with monolingual data. While this technique has been previously investigated in the study of natural language processing (Chi et al., 2019; Aggarwal et al., 2021), we are not aware of any prior work in the Digital Humanities explicitly employing a multilingual classifier trained solely on monolingual data. The consequences of this technique are innumerable for low-resource languages whose digitized material may not be sufficient in volume to train bespoke models. Future researchers will want to verify whether the same technique can be applied in other classification tasks (i.e. topic classification, sentiment analysis).

We release together with the dataset a collection of Python scripts intended to enable researchers to more easily access and conduct experiments on private works accessible on the HTRC Data Capsule at https://git.sr.ht/~srhm/hathi-scripts. While HathiTrust does provide a set of basic utilities for conducting certain NLP experiments, downloading larger volumes of text remains a disproportionately difficult task especially given the security conditions of the capsule. We hope these scripts will aid future researchers in conducting experiments on the platform.

REPOSITORY LOCATION

https://doi.org/10.6084/m9.figshare.21354798.

FUNDING INFORMATION

This work is supported by funding from the Social Sciences and Humanities Research Council of Canada (895-2013-1011).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Sil Hamilton: Methodology, Investigation, Writing

Andrew Piper: Conceptualization, Resources, Supervision

AUTHOR AFFILIATIONS

Sil Hamilton orcid.org/0000-0002-6579-4628
Languages, Literatures, and Cultures, McGill University, Montreal, Canada

Andrew Piper orcid.org/0000-0001-9663-5999
Languages, Literatures, and Cultures, McGill University, Montreal, Canada

REFERENCES

Aggarwal, S., Kumar, S., & **Mamidi, R.** (2021). Efficient multilingual text classification for Indian languages. *Proceedings of Recent Advances in Natural Language Processing* (pp. 19–25). DOI: https://doi.org/10.26615/978-954-452-072-4_003

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. DOI: https://doi.org/10.5334/johd.71

Chi, Z., Dong, L., Wei, F., Mao, X., & **Huang, H.** (2019). Can monolingual pretrained models help cross-Lingual classification? *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. DOI: https://doi.org/10.48550/arXiv.1911.03913

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., & **Stoyanov, V.** (2019). Unsupervised cross-lingual representation learning at scale. *Proceeding of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). DOI: https://doi.org/10.18653/v1/2020.acl-main.747

Evans, E., & **Wilkens, M.** (2018). Nation, Ethnicity, and the Geography of British Fiction, 1880–1940. *Journal of Cultural Analytics*, 3(2). DOI: https://doi.org/10.22148/16.024

Fenlon, K., Fallaw, C., Cole, T., & **Han, M.** (2014). A preliminary evaluation of HathiTrust metadata: Assessing the sufficiency of legacy records. *IEEE/ACM Joint Conference on Digital Libraries* (pp. 317–320). DOI: https://doi.org/10.1109/JCDL.2014.6970186

Gil, A., & **Ortega, É.** (2016). Global outlooks in digital humanities: Multilingual practices and minimal computing. In C. Crompton, R. Lane & R. Siemens (Eds.), *Doing Digital Humanities: Practice, training, research* (pp. 22–34). London; New York: Routledge.

Mahony, S. (2018). Cultural diversity and the Digital Humanities. *Fudan Journal of the Humanities and Social Sciences*, *11*(3), 371–388. DOI: https://doi.org/10.1007/s40647-018-0216-0

Odebrecht, C., Burnard, L., & **Schöch, C.** (2021). European Literary Text Collection (ELTeC): April 2021 release with 14 collections of at least 50 novels. (v1.1.0). *Zenodo*. DOI: https://doi.org/10.5281/zenodo.4662444

ONNX Runtime developers. (2021). *ONNX Runtime*. https://onnxruntime.ai.

Piper, A. (2022). The CONLIT dataset of contemporary literature. *Journal of Open Humanities Data*, *8*, 24. DOI: https://doi.org/10.5334/johd.88

Ryan, M.-L. (1980). Fiction, non-factuals, and the principle of minimal departure. *Poetics*, *9*(4), 403–422. DOI: https://doi.org/10.1016/0304-422X(80)90030-3

Spence, P. J., & **Brandao, R.** (2021). Towards language sensitivity and diversity in the digital humanities. *Digital Studies/Le champ numérique*, *11*(1). DOI: https://doi.org/10.16995/dscn.8098

Underwood, T. (2014). Understanding genre in a collection of a million volumes. *Interim Performance Report for the Digital Humanities Start-up Grant*. DOI: https://doi.org/10.17613/M6W07V

Van Eijnatten, J., Pieters, T., & **Verheul, J.** (2013). Big Data for global history: The transformative promise of Digital Humanities. *Low Countries Historical Review*, *128*(4), 55–77. DOI: https://doi.org/10.18352/bmgn-lchr.9350

TO CITE THIS ARTICLE:
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, pp. 1–7. DOI: https://doi.org/10.5334/johd.95

Published: 08 February 2023

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