Journal of Open
Humanities Data

The TRANSCOMP Dataset of Literary Translations from 120 Languages and a Parallel Collection of English-language Originals

DATA PAPER

MATT ERLIN
ANDREW PIPER
DOUGLAS KNOX
STEPHEN PENTECOST
ALLIE BLANK

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

]u[ubiquity press

ABSTRACT

The TRANSCOMP Dataset of Literary Translations is a collection of document-level word frequencies sampled from 10,631 translations into English of global literary fiction published since 1950, together with a historically matched parallel corpus of 10,682 fictional works originally published in English. We provide CSV files with word frequency counts for 10,000-word samples taken from each text. The associated metadata is available in a separate CSV. These data will be useful to literary scholars and linguists working in translation studies, and those interested in the linguistic, stylistic, and thematic specificity of translations from particular regions.

CORRESPONDING AUTHOR:

Matt Erlin

Germanic Languages and Literatures, Washington University, St. Louis, US

merlin@wustl.edu

KEYWORDS:
translation studies; computational literary studies; world literature; natural language processing; text corpus; text collection

TO CITE THIS ARTICLE:
Erlin, M., Piper, A., Knox, D., Pentecost, S., & Blank, A. (2022). The TRANSCOMP Dataset of Literary Translations from 120 Languages and a Parallel Collection of English-language Originals. *Journal of Open Humanities Data*, 8: 29, pp. 1–6. DOI: https://doi.org/10.5334/johd.94

(1) OVERVIEW

REPOSITORY LOCATION

doi.org/10.7910/DVN/ITLGQV

CONTEXT

This dataset consists of document-level word frequency samples drawn from a parallel corpus containing 10,631 translations of literary fiction into English from 120 different languages published since 1950 along with a comparable set of 10,682 works of fiction written originally in English during the same time period. All texts are contained in the Hathi Trust Digital Library and are derived from the ca. 176,000-volume NovelTM collection created by Underwood et al. (2020).

The dataset was compiled as part of an ongoing research project into the unique linguistic, stylistic, and thematic features of translated fiction as compared to fiction written originally in English. Following the precedent established by Toury's (1980) and Baker's (1993) pioneering work on translation universals, our aim has been to create two independent corpora that enable researchers to evaluate translated texts as they relate to target language texts *in general,* rather than to compile a corpus of *translations* and their corresponding *source texts.* While corpora designed for comparative translation studies do exist, including a number of parallel corpora, they are often focused on single pairs of languages and/or non-literary texts; moreover, they are not constructed to facilitate the kind of historical comparisons that interest computational literary scholars. To our knowledge, no existing collection of historically matched translated and original-language fictional texts even approaches the size or linguistic diversity of our corpus, and we hope that it will serve as a resource for additional research.

(2) METHOD

STEPS

On the basis of the metadata provided by Underwood et al. (2020) regarding the NovelTM dataset of English-language fiction, we first used a set of regular expressions such as "translated from the Swedish," "from the [language]," "tr. from," "rendered into English," etc. to identify an initial list of translated texts. Next, if an author was included in this initial list, we included all titles by that author. For example, if one volume by Leo Tolstoy had "translated from" in one of its metadata fields, we included all works by Leo Tolstoy in our set of translations.

Original English-language works were identified by fuzzy matching against a large set of author names derived from Wikipedia and the Virtual International Authority File (VIAF), which consists of millions author names derived from 68 library catalogues from around the world. Any names identified as English-language authors from this list were then removed from the translation data. We similarly used non-English-language author data to match with our translation data and reviewed all non-matching works by hand.

Information on a translation's original language was taken from two primary sources: explicit references included in the titles of the works in Hathi (e.g., translated from the Swedish) and from the HathiTrust extracted features metadata. These results were supplemented using fuzzy matching of author lists from Wikipedia and VIAF. The remaining missing data was manually retrieved using WorldCat and other internet sources. To identify date of publication, we used Underwood et al.'s "inferred date" (2020). Because the holdings of translations in Hathi are heavily skewed toward a rather small set of European authors and languages in the first part of the twentieth century, we subsetted our data down to the date range 1950–2008, which aligns with the period construct of "post-war" fiction used in literary studies (McGurl, 2009). Finally, we also removed all volumes where Underwood et al.'s predicted probability of being non-fiction was greater than 85% (2020). Given that the set of original language works was larger than the set of translations, we also randomly downsampled each year of our original publications to match the number of translations.

We then processed the files to be extracted as word frequency data. Working within the HTRC capsule (Plale et al., 2019), we first downloaded individual page files using the preloaded functions in the HTRC Workset Toolkit to remove running page headers and footers. For each volume, we concatenated individual page files into a single document. After tokenizing with regular expressions, we next represented each document as ten randomly selected 1,000-continuousword samples drawn from the middle 60% of the document to avoid paratextual content in the front and back matter. This sampling enables us to control for effects that might arise from the different lengths of the source texts. To mitigate problems related to low OCR quality, foreign-language passages, or the presence of other non-standard characters, only samples that had 90% of words in an English dictionary were kept. If a work did not have ten samples that met this criteria, it was removed. All of this work was completed in the Hathi capsule. These samples were then converted into bags of words, which we are able to make accessible to the scholarly community in the form of two CSV files, one for originals and one for translations, listing raw frequency counts by document for each of the words in each of the original document samples.

While the final corpus of translated texts remains skewed towards European languages, it does include a significant number of works originally published in East Asian and South Asian languages and a smaller number of works originally published in Middle Eastern and African languages. Figures 1–4 provide an overview of the dataset.

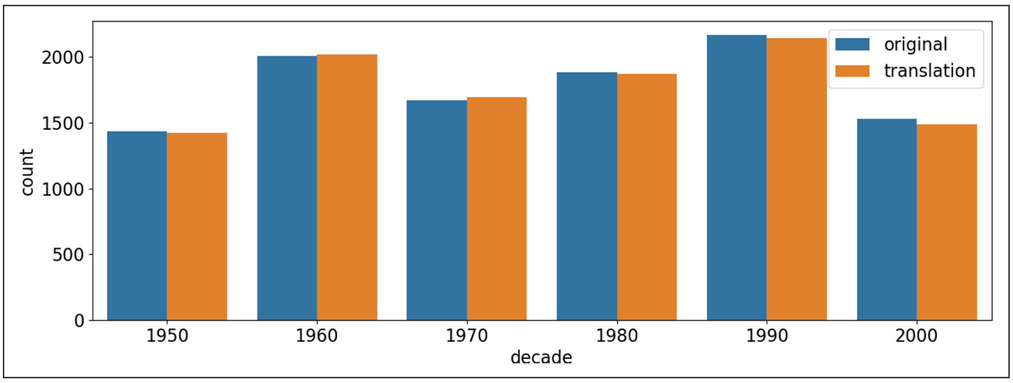

Figure 1 Count of works by decade, originals and translations.

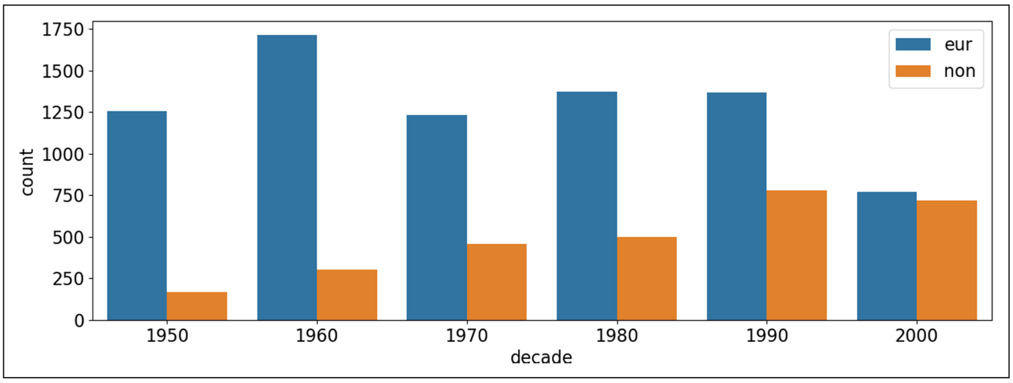

Figure 2 Count of translated works by decade, non-European and European.

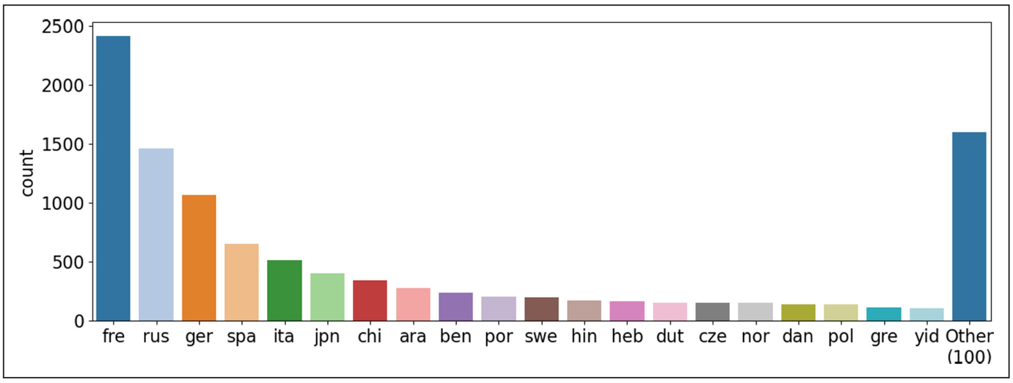

Figure 3 Count of translations from the top 20 languages represented in the corpus.

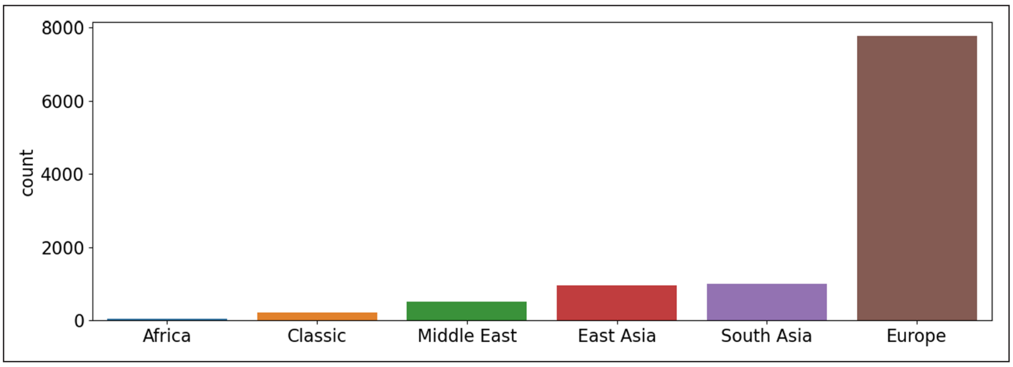

Figure 4 Total translations by subregion (classical literature as separate category).

QUALITY CONTROL

To test the accuracy of our identification of translations in the NovelTM dataset, we created a random sample of 100 works identified as translations and 100 works identified as originals from our data. We then manually checked each title to see whether our classification had been correct. We found that 99 were correctly labeled for an estimated precision of .99. We did not evaluate the accuracy of recall (i.e., translations in Hathi that we missed). In addition to its impracticality, given the size of the original dataset, the results would simply have told us whether our sample was representative of translations in the Hathi corpus. For the comparative work we envision, the key question is whether we have a randomly sampled set of translations that mirrors our English original corpus, not whether it accurately represents the distribution of texts in Hathi.

LIMITATIONS

One key limitation is the date range of our data (1950–2000). Expanding this date range, however, leads to an overwhelming predominance of a few European languages, which runs counter to our goal of having a diverse set of source languages represented. As Figure 2 reveals, even after 1950, translations in the Hathi Library skew European. Whether this is true of the English-language market for fiction more generally or is an artifact of Hathi we leave for future work. We note that the period 1950–2000 is considered a distinct period within literary history and thus our data aligns with this historical construct (McGurl, 2009).

An additional potential limitation is the presence of works in the dataset that were originally published prior to 1950 but which were translated or re-translated at a later date. On the basis of the (incomplete) information that we have on author birth and death dates, we estimate that such works constitute between 15–20% of the total (see Python notebook in the repository). Finally, our data is limited due to intellectual property restrictions that only allow us to export word frequencies and not the full text from Hathi. We provide all Hathi IDs such that researchers can recreate our data inside of the Hathi capsule system.

(3) DATASET DESCRIPTION

OBJECT NAME

The dataset consists of three CSV files: Translation_samples.csv, Original_samples.csv, and TransComp_metadata.csv. We have also included a Python notebook addressing the question of original publication dates: 1950_boundary_question.ipynb

FORMAT NAMES AND VERSIONS

CSV, ipynb

CREATION DATES

2021–04–28 – 2022–04–18

DATASET CREATORS

Allie Blank

Douglas Knox

Stephen Pentecost

LANGUAGE

English

LICENSE

CC0

REPOSITORY NAME

Dataverse

PUBLICATION DATE

2022–10–07

(4) REUSE POTENTIAL

Two primary areas of research will likely benefit from access to this data. Recent scholarship in the sociology of translation (Heilbron, 1999; Bachleitner and Wolf, 2004; Sapiro, 2016; 2020) has helped reveal the structural asymmetries in the global flow of translations, often adopting a core—semi-periphery—periphery model to clarify the dominant role played by a small subset of European languages in this regard. To date, however, there has been virtually no effort to link these asymmetries to differences in the linguistic, stylistical, or thematic features of translations (Piper and Erlin 2022). To what extent, in other words, do translations from "peripheral" languages or language regions exhibit common features that might reinforce or challenge existing cultural biases or reflect the pressures imposed on "peripheral" authors in what Pascale Casanova (2004) has referred to as the "world republic of letters"? We believe that our data set will greatly facilitate the investigation of such questions.

While only having bags of word frequencies places some limitations on what is possible in this regard, prior research has generated important cultural insights using such word distribution approaches (Erlin, 2017; Jockers and Mimno, 2013; Piper, 2016; Underwood, 2016). In addition, the CSV files include information on the page count for each work sampled as well as the mean sentence length for the samples, the latter of which we calculated in the Hathi capsule. Finally, we include metadata so that researchers can work on the full texts within the Hathi data capsule system.

With regard to translation studies more broadly, we believe that this historically matched collection of translations and originals can shed new light on questions of "translationese" (i.e. translation universals). Corpus and computational linguists have long been identifying ways in which translation can be thought of as a distinct linguistic practice that consists of quasi-universal behaviors conditioned by the nature of moving between languages and the cognitive demands of doing so (Volansky, Orden, and Winter, 2015). Only a few studies, however, have focused on the specific qualities of literary translations, and certainly not at the scale made possible by this dataset. We think the collection is particularly well suited to investigations into the question of whether translations can be understood as a literary *genre* (Piper and Erlin, 2022). While the concept of genre is famously multivalent in literary studies (Cohen, 2017, 86), we use the term in the most elementary sense as a set of works that exhibit "shared features" (Reichert, 1978, 57) – translations in this case — that can be algorithmically classified on the basis of its relational distinctiveness vis-a-vis non-translated works as well as the ways it coheres as a category over time.

ADDITIONAL FILE

The additional file for this article can be found as follows:

- **Supplementary Material.** About the TRANSCOMP dataset. DOI: https://doi.org/10.5334/johd.94.s1

ACKNOWLEDGEMENTS

We are grateful to HathiTrust for the permission to release this data.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

- Matt Erlin: conceptualization, methodology, writing, visualization
- Andrew Piper: conceptualization, methodology, writing, visualization

- Douglas Knox: conceptualization, methodology, writing, data curation, visualization
- Stephen Pentecost: conceptualization, methodology, data curation
- Allie Blank: conceptualization, methodology, data curation

AUTHOR AFFILIATIONS

Matt Erlin orcid.org/0000-0002-0536-7499
Germanic Languages and Literatures, Washington University, St. Louis, US
Andrew Piper orcid.org/0000-0001-9663-5999
Languages, Literatures, and Cultures, McGill University, Montreal, Canada
Douglas Knox orcid.org/0000-0002-7168-7271
Humanities Digital Workshop, Washington University, St. Louis, US
Stephen Pentecost orcid.org/0000-0002-2093-6151
Humanities Digital Workshop, Washington University, St. Louis, US
Allie Blank
Humanities Digital Workshop, Washington University, St. Louis, US

REFERENCES

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In Baker, M., Francis, G., Tognini-Bonelli, E. (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233–250). Amsterdam/Philadelphia: Benjamins. DOI: https://doi.org/10.1075/z.64.15bak

Bachleitner, N., & **Wolf, M.** (2004). Auf dem Weg zu einer Soziologie der literarischen Übersetzung im deutschsprachigen Raum. *Internationales Archiv Für Sozialgeschichte Der Deutschen Literatur, 29*(2), 1–25. DOI: https://doi.org/10.1515/IASL.2004.2.1

Casanova, P. (2004). *The world republic of letters*. Cambridge: Harvard University Press.

Cohen, R. (2017). *Genre theory and historical change: Theoretical essays of Ralph Cohen*. Charlottesville: University of Virginia Press.

Erlin, M. (2017). Topic modeling, epistemology, and the English and German novel. *Journal of Cultural Analytics*, *2*(2), 11070. DOI: https://doi.org/10.22148/16.014

Heilbron, J. (1999). Towards a sociology of translation: Book translations as a cultural world-system. *European Journal of Social Theory*, *2*(4), 429–444. DOI: https://doi.org/10.1177/136843199002004002

Jockers, M. L., & **Mimno, D.** (2013). Significant themes in 19th-century literature. *Poetics*, *41*(6), 750–769. DOI: https://doi.org/10.1016/j.poetic.2013.08.005

McGurl, M. (2009). *The program era: Postwar fiction and the rise of creative writing*. Cambridge: Harvard University Press. DOI: https://doi.org/10.2307/j.ctvjsf59f

Piper, A. (2016). Fictionality. *Journal of Cultural Analytics, 2*(2). DOI: https://doi.org/10.22148/16.011

Piper, A., & **Erlin, M.** (2022). The predictability of literary translation. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pp. 155–160.

Plale, B., Dickson, E., Kouper, I., Liyanage, S. H., Ma, Y., McDonald, R. H., Walsh, J. A., & **Withana, S.** (2019). Safe open science for restricted data. *Data and Information Management*, *3*(1). DOI: https://doi.org/10.2478/dim-2019-0005

Reichert, J. (1978). More than kin and less than kind: The limits of genre theory. In J. P. Strelka (Ed.), *Theories of Literary Genre*, pp. 57–79. University Park: Pennsylvania State University Press.

Sapiro, G. (2016). How do literary works cross borders (or not)?: A sociological approach to world literature. *Journal of World Literature*, *1*(1), 81–96. DOI: https://doi.org/10.1163/24056480-00101009

Sapiro, G. (2020). The transnational literary field between (inter)-nationalism and cosmopolitanism. *Journal of World Literature*, *5*(4), 481–504. DOI: https://doi.org/10.1163/24056480-00504002

Toury, G. (1980). *In search of a theory of translation*. Tel Aviv: Porter Institute for Poetics and Semiotics, Tel Aviv University.

Underwood, T., Kimutis, P., & **Witte, J.** (2020). NovelTM datasets for English-language fiction, 1700–2009. *Journal of Cultural Analytics*, *5*(2), 13147. DOI: https://doi.org/10.22148/001c.13147

Underwood, T. (2016). The Life Cycles of Genres. *Journal of Cultural Analytics, 2*(2). DOI: https://doi.org/10.22148/16.005

Volansky, V., Ordan, N., & **Wintner, S.** (2015). On the features of translationese. *Digital Scholarship in the Humanities*, *30*(1), 98–118. DOI: https://doi.org/10.1093/llc/fqt031

TO CITE THIS ARTICLE:
Erlin, M., Piper, A., Knox, D., Pentecost, S., & Blank, A. (2022). The TRANSCOMP Dataset of Literary Translations from 120 Languages and a Parallel Collection of English-language Originals. *Journal of Open Humanities Data*, 8: 29, pp. 1–6. DOI: https://doi.org/10.5334/johd.94

Published: 26 December 2022