

ARTICLE

Biodiversity is not declining in fiction

Andrew Piper¹ ¹ McGill University

Keywords: environmental humanities, fiction, English literature, replication, biodiversity

<https://doi.org/10.22148/001c.38739>

Journal of Cultural AnalyticsVol. 7, Issue 3, 2022

This paper attempts to replicate the findings of the recent work, “The rise and fall of biodiversity in literature,” by Langer et al. (2021). Using a large corpus from Project Gutenberg (N = ~15,000) and a dictionary-matching method of over 240K biological taxa, Langer et al. find that the frequency and diversity of biological taxa have been declining steadily since the first half of the nineteenth century, echoing prior work in cultural analytics. This paper applies the original paper’s three primary measures to two additional data sets along with the original dataset and compares their dictionary-based method with an alternative supervised machine learning method. I find that the trajectory of biological tokens in fiction in the new data sets is directionally opposite to that shown by Langer et al. independent of the methods used (i.e. taxa rise rather than fall since the first half of the nineteenth century) but that their breakpoint estimation appears largely robust within +/- 15 years. Based on this analysis, I suggest that the discrepancy between our results is due to corpus construction rather than choice of method. I find that only conditioning on fiction in the original dataset generates results more similar to the two alternative datasets used here. In addition to emphasizing the importance of corpus construction for cultural analytics, these findings also raise larger questions about the difficulties of interpreting lexical items as indices of social attitudes, pointing to a need for future work.

Introduction

In a recent important work, Langer et al. (2021) measure the prevalence and diversity of biological entities in creative literature over the past three centuries. Building off of prior work (Kesebir and Kesebir; Travis), they are interested in better understanding humans’ relationship to nature through the imaginative depiction of nature over time. As they trenchantly point out, “There is still a fundamental lack of understanding regarding the influence of nature on various aspects of our culture and its development” (Langer et al. 1094). The large-scale study of cultural attitudes towards nature across long historical periods of human history is indeed an important area of study and in need of further work (Travis; Lee and Beckelhimer).

Building a database of over 200,000 biological entities (“taxon labels”) and working with a Project Gutenberg corpus of ~15,000 books, they find that the frequency and diversity of biological taxa have been declining steadily since the first half of the nineteenth century, echoing prior work (Kesebir and Kesebir). As the authors write, “We show that richness, abundance and Shannon diversity peak in the 1830s, followed by a consistent decline over more than 100 years until the middle of the 20th century” (Langer et al. 1101).

As those working in the field of cultural analytics are aware, creating an adequate representation of historical classes of writing has proven challenging (Bode; Schmidt et al.). Different sampling methods and different classification methods may generate meaningfully different results. At the same time, the estimation of linguistic categories like “biological taxa” is also not straightforward. Pursuant to the authors’ point that nature suffuses cultural practices (and vice versa), some meaningful portion of human language borrows natural terms to refer to cultural entities and vice versa. Proper names like “Rose” and “Iris” are two obvious examples of the former practice, just as the label “Queen Bee” points to the reverse process of using cultural markers to name natural entities (which are then recycled to refer to human stereotypes). “References” to biological taxa may thus encompass *indexical* references to real-world objects (as when a writer talks about a particular kind of tree) or *residual* references where natural words are used to refer to other kinds of entities (as in proper names or derogatory terms like “he’s such a leech”). As critics have long pointed out and Langer et al. underscore, the nature-culture divide is not easily drawn.

Because of these difficulties, and because of the intrinsic value of their project, it is worth testing the extent to which their principal findings are reproducible using different data sets and different methods. As with all efforts at replication, no single project can either confirm or refute initial findings, but they can begin to give us confidence about the robustness of prior insights, or, in the case that the findings are not reproducible, point us in new directions.

This paper compares the authors’ methods of dictionary-based taxa estimation with an alternative method based on supervised machine learning. Both methods are applied to the original corpus as well as two alternative data sets: the first is a large scale collection of ~1.6 million pages of English prose sampled from over 300,000 books published since 1800 contained in the Hathi Trust Digital Library (Bagga and Piper) and the second a considerably smaller collection (~1,700 volumes) of English-language fiction published over the same time period (Piper, *Enumerations*). Based on these comparative methods and samples, this paper shows that:

- the trajectory of biological tokens in fiction is directionally opposite to that shown by Langer et al. independent of the dataset or methods used (i.e. in all cases taxa rise rather than fall since the first half of the nineteenth century);
- the trajectory of biological tokens in non-fiction is even more strongly downward oriented than the authors report suggesting that one of the discrepancies between our results may likely be due to their combination of fiction with non-fiction under their larger heading “creative literature”;

- the authors' breakpoint estimation of meaningful change occurring around the period of 1835 appears to be robust within +/- 15 years for fiction for the two new datasets;
- the database matching method used by the authors produces high levels of false positives leading to increased-estimates of the prevalence of taxa labels by roughly 80%;
- nevertheless there remains a very high correlation between the machine-learning and dictionary-based methods ($r > 0.93$), a surprising finding in its own right, but one that suggests that the discrepancy between results is far more sample- than method-driven;

Given these findings, which will be discussed in greater detail here, two key insights emerge. The first is that findings in cultural analytics can be highly sensitive to corpus construction (Bode; Schmidt et al.). In particular, as prior work has shown both "genre" (Argamon et al.; Piper and Portelance; Wilkens) and "instrumentality," i.e. fiction/non-fiction (Piper, "Fictionality"; Sap et al.), have strong effects on the stylistic behavior of authors (Underwood, "Genre Theory and Historicism"). As I show, disaggregating data along these lines can result in directionally opposing historical behavior and thus needs to be taken into consideration whenever cultural corpora are constructed and analyzed for historical processes.

A second more speculative insight that may be on display here is the way the distribution of biological entities in fiction might represent something other than a commitment to or strict representation of the natural world. Langer et al. write, "As texts are cultural products of their time, we may assume that the usage of taxon labels in those texts is correlated with the societal awareness of biodiversity at that time." But what the data may suggest, at least for fiction, is that given both the highly generic nature of taxa and the dramatic shift ca. 1850 followed by consistent historical behavior taxon labels may not *primarily* be representations of "social attitudes" towards biodiversity. Rather, they may function as indices of some larger generic developments in the domain of creative writing, where biological entities are one component of a larger semantic shift. As I discuss in the conclusion, understanding this problem will take further research and represents a significant program of research for the field of cultural analytics.

Data

Three principal data sets are used for this paper. The first is based on the original paper's use of Project Gutenberg (hereafter PG). From the over 59,000 volumes in PG, Langer et al. select 15,798 based on criteria listed in their paper, which they refer to as "creative literature." From these, they keep only those where the author's birth and death date are available, which Langer et al. use to estimate date of publication (which is not included in PG metadata). This

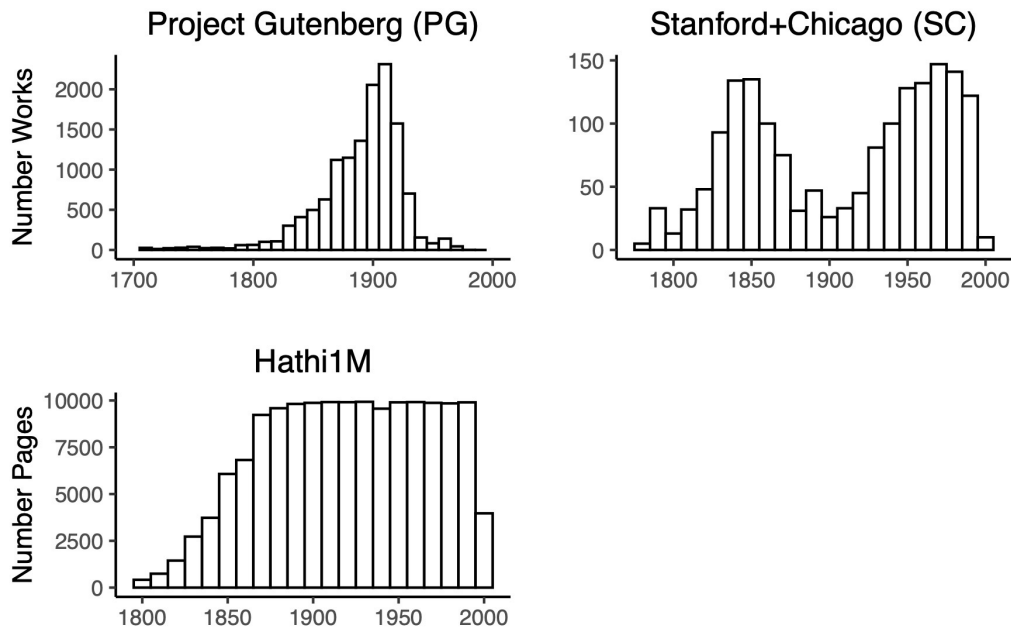


Figure 1. Distribution of works and pages over time for the three data sets.

results in 15,238 books. Books are then filtered by requiring a minimum of 15,000 tokens as indicated by Langer et al. in their paper resulting in 13,075 books, a notably smaller number than that reported by Langer et al.

The second data set is the Hathi1M dataset, which consists of 1,671,370 pages of English-language prose drawn from over 300,000 volumes from the Hathi Trust Digital Library spanning the previous two-hundred years (Bagga and Piper). Pages in this dataset have been classified as either belonging to works of fiction or non-fiction using a single predictive model based on prior work (Underwood et al.).

The third dataset, known as the Stanford+Chicago collection (hereafter SC), consists of 1,711 works of fiction published during the same time period and is more fully described in Piper (*Enumerations*). It comprises a combination of the Chadwyck Healey database of nineteenth-century English-language fiction curated by the Stanford Literary Lab plus works of English-language fiction published since 1880 made available through Amazon curated by the Chicago Textlab. The sample used here represents a random sample of 1,000 works drawn from the larger Chicago collection and 711 works from the Chadwyck Healey collection. [Figure 1](#) provides a distribution of pages/works over time for the three datasets and [Table 1](#) lists the limitations and affordances of each data set.

Methods

This paper compares two methods to predict the presence of biological taxa in texts. The first is based on the methods of Langer et al. which I label the “database method” (hereafter DB method). Langer et al. develop a large

Table 1. List of affordances and limitations of each data set

Dataset Name	Total Tokens	Affordances	Limitations
PG 12,574 volumes	1,140,676,329	<ul style="list-style-type: none"> - large collection of open-access full text data - non-OCR 	<ul style="list-style-type: none"> - does not continue through end of twentieth-century - non-uniform distribution over time - sampling criteria not well known - no genre differentiation - sparse pre-1800 and post-1940
SC 1,711 volumes	217,854,521	<ul style="list-style-type: none"> - manually curated - contains only fiction - non-OCR - continuous from 1800 to 2000 	<ul style="list-style-type: none"> - small size to measure historical change - non-uniform distribution over time - 19C and 20C sampling criteria not uniform
Hathi1M 1,671,370 pages	587,951,218	<ul style="list-style-type: none"> - large historically diverse sample - uniformly distributed over time - continuous from 1800 to 2000 - differentiated by instrumentality, i.e. fiction and non-fiction 	<ul style="list-style-type: none"> - some error due to OCR - only page-level derived data - labels are based on predictive models not manual curation and thus contain some error

database of 242,443 taxa labels derived from Wikipedia along with a small list of blackout-tokens to be removed (such as “bishop,” “red,” etc.). Word tokens are lemmatized and then matched against the database. While not reported in their paper, verbal communication indicated that only lemmas were kept that were tagged as nouns using a part-of-speech tagger.

The second method is based on a supervised model using machine learning (hereafter ML method) that predicts the presence of biological taxa given a word’s context. For this method, the “supersense” annotations as implemented in bookNLP are used, which utilizes the Wordnet taxonomy and makes predictions about a word’s category based on its local context. The entities “noun.plant” and “noun.animal” are used to estimate the presence of biological taxa in a given text.

All data processing for this paper was done using David Bamman’s bookNLP (Bamman et al.), which tokenizes, lemmatizes, tags for part-of-speech and dependency relationships and provides the above “supersense” annotations.

The problems of polysemy facing dictionary-matching methods are well-known. For example, false-positives can occur when words included in the taxa list are used to refer to non-biological entities, as when “Rose” refers to a character rather than a plant or “bear” refers to an action not an animal. As mentioned above, the first instance might still be considered a biological reference in a broader sense, but the second example definitely not. In some cases, spurious entries may also pose a problem, as in this case with tokens like “end” or “beauty.” False negatives occur when the lexicon or database is missing terms important to the domain.

On the other hand, predictive models may fail to capture a word’s true sense given insufficient training, complex sentence structure, or novel domains. For example, the categories of “plant” and “animal” may not account for all types of mentions of biological taxa in fiction, nor will they count residual references

Table 2. Validation of the two methods used in this paper.

Method	Precision	Recall	F1
DB	0.733	0.8	0.765
ML	0.797	0.89	0.841

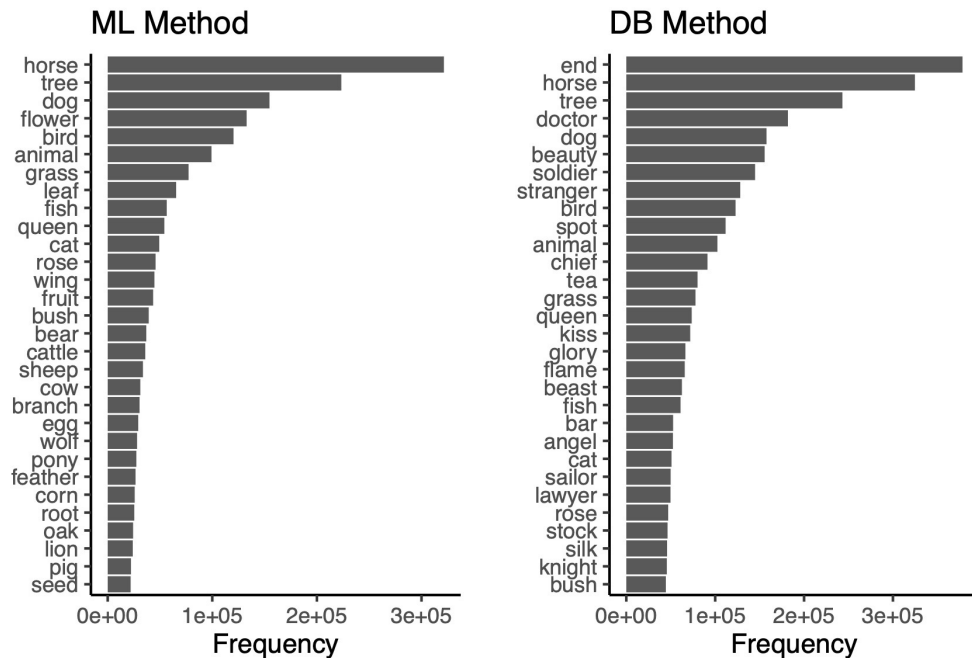


Figure 2. Lists of most frequent tokens identified by the two methods using the PG dataset

like proper names. To clarify the efficacy of the two methods, I validate them on a small sample of one hundred 50-token passages (i.e. 5,000 tokens). Because biological taxa are rare linguistic events the sample conditions on passages where at least one token has been predicted to be a taxon. As can be seen in [Table 2](#), the machine learning method outperforms the DB method by about 7.5 points, though I expect that this number would be higher given a truly random sample where the precision of the DB method would likely decrease. It is worth noting that not using the part-of-speech tagging condition of nouns-only further depresses the accuracy of the database method by several more points.

As a way of further validating the two methods, the list of most frequent taxa identified by each method are presented similar to Langer et al. ([Fig. 2](#)). One can see here the higher numbers of false positives in the DB method (“end”, “doctor”, “soldier”, “stranger”, etc.). These words do not appear in Langer et al.’s presentation of most frequent taxa, suggesting that there may have been an additional cleaning step of the data after implementation of the database.

In addition to the identification of biological taxa within texts, Langer et al. measure the prevalence of taxa according to three frameworks. [Table 3](#) describes the implementation of their three primary measures used in this

Table 3. Description of the three primary measures used in Langer et al. and the implementation in this paper.

Category	Measure	Implementation
Abundance	Token frequency	# taxa / total words per book / page
Richness	Normalized type frequency	mean # types per ten 1,000 word samples per book
Diversity	Normalized type entropy	mean entropy of types per ten 1,000 word samples per book

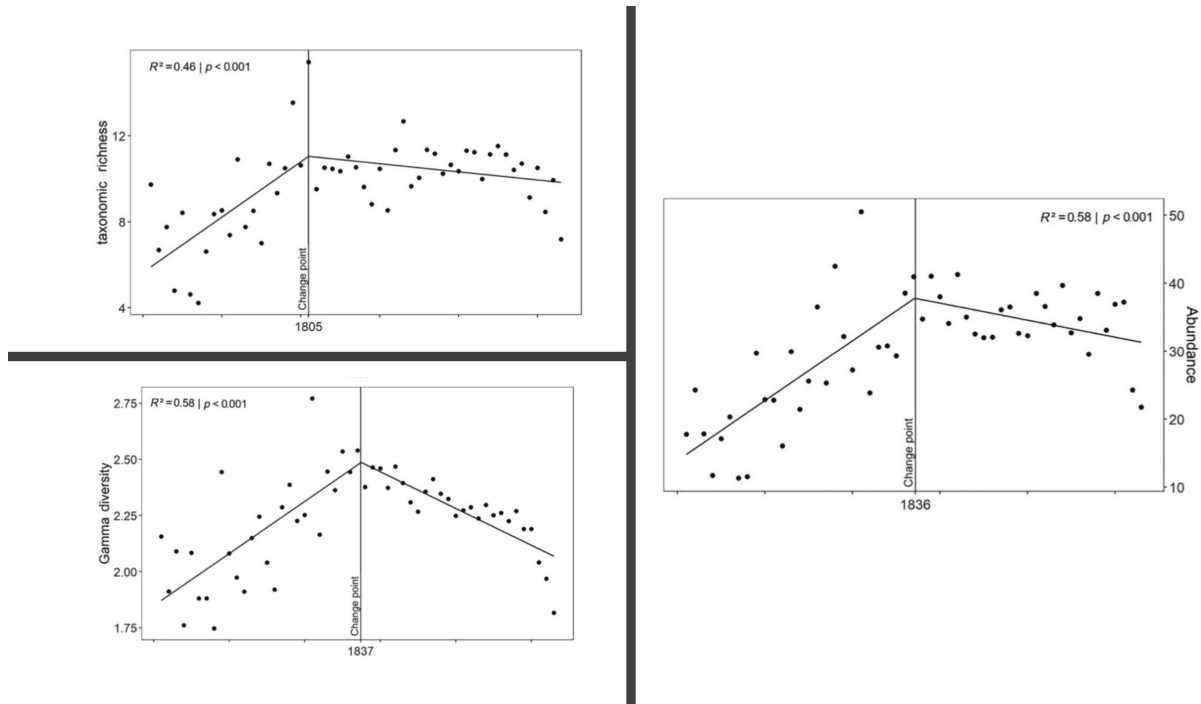


Figure 3. The three primary measures presented in Langer et al. using PG data

paper. As in Langer et al., each measure is averaged over five-year periods to derive final estimates. Boot-strap sampling is also used for the PG dataset to arrive at five-year estimates as indicated by Langer et al.

Results

The principal aim of Langer et al.'s paper is to estimate the historical trajectory of the representation of biological taxa in English-language creative writing. They do so by aggregating their data into 5-year periods and then estimating a breakpoint across their historical time period. Their three main graphs are provided ([Fig. 3](#)) and then my results ([Fig. 4-6](#)). Because the Hathi1M data only provides derived supersense data, I only implement the ML method on this data and only for the Abundance measure. However, I break out fiction and non-fiction separately to illustrate the effects of instrumentality on the estimated prevalence of taxa. I also subset the PG data by only keeping works that have “fiction” in their subject column to observe what effect this has on the results.

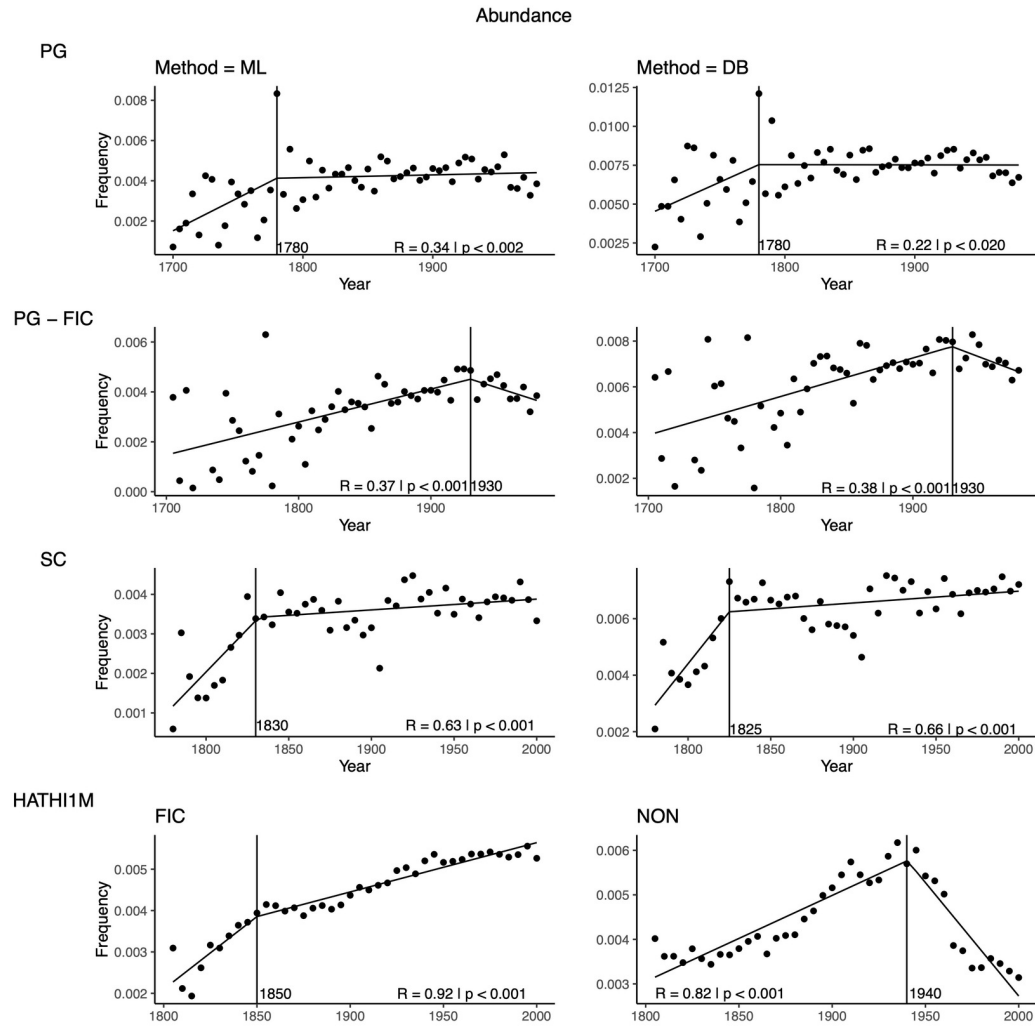


Figure 4. Abundance measures in different data sets using the two principal methods.

Rows 1-3 are the DB and ML methods applied to the original PG dataset, the PG dataset with only fiction, and the SC dataset respectively. The fourth row represents only the ML method applied to the Hathi1M data disaggregated by instrumentality (fiction / non-fiction).

As we can see in Row 1 of [Figs. 4-6](#), the implementation of the DB method applied to the PG dataset partially replicates the results of Langer et al. The estimated breakpoints are consistently earlier, but I either find no rise in the number of estimated taxa or a decline of diversity similar to the original paper.

However, when we look at the fiction-only subset of the PG data along with the two alternative datasets (Rows 2-4), there is a notable *increased* rate of taxa abundance and taxa diversity after the original paper's estimated breakpoint of ~1835 for both methods. In other words, regardless of method used when one conditions on datasets only comprised of fiction we see directionally opposite behavior to that reported in the original paper. Also worth noting is that while the original paper emphasized the centrality of “diversity” – that there is a cultural significance to not simply using fewer taxa but fewer *kinds* of taxa – these measures (abundance and richness) actually correlate extremely highly ($r = 0.89$, see [Fig. 7](#)). Finally, I also use one of the affordances of the ML method, which can disambiguate between plant and animal taxa, to illustrate

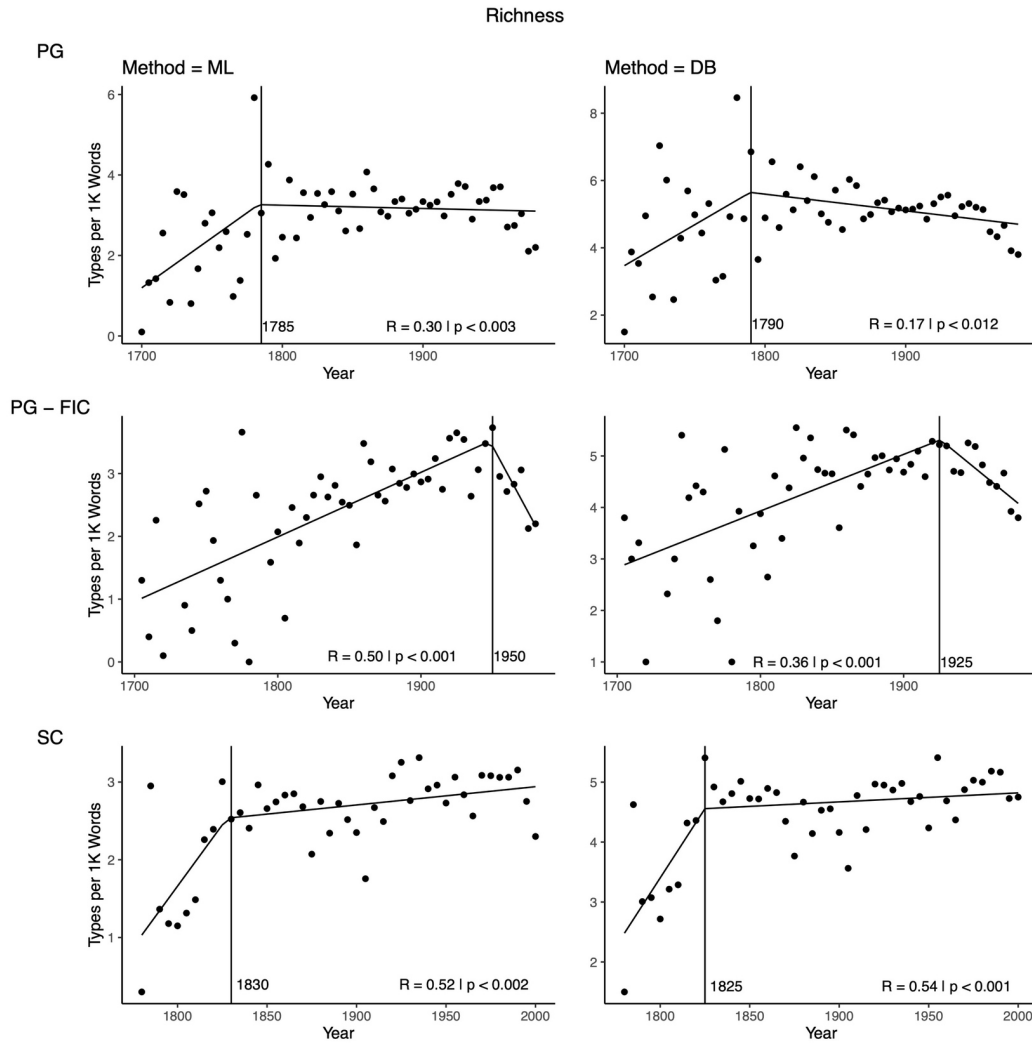


Figure 5. Richness measures comparing the two methods.

how fictional narratives are consistently more strongly weighted towards the representation of animals than plants, maintaining a biological hierarchy of various life forms that appears to be invariant across time (Fig. 8).

Discussion

The aim of this paper has been to test Langer et al.'s findings concerning the decline of biodiversity after the mid-1830s in English-language creative literature. To do so, I have introduced two alternative datasets as well as one alternative method. The machine learning based method is demonstrably more reliable for the prediction of biological taxa and it is recommended that future work utilize similar methods for the estimation of linguistic types where possible.

Second, and more significantly, the effects observed by Langer et al. are reversed when conditioning only on subsets of “fiction” within the category of creative writing. References to biological taxa – at least with respect to plant and animal life – appear to have been steadily increasing over time in imaginary prose storytelling. I surmise that the downward trend observed by Langer et al. is

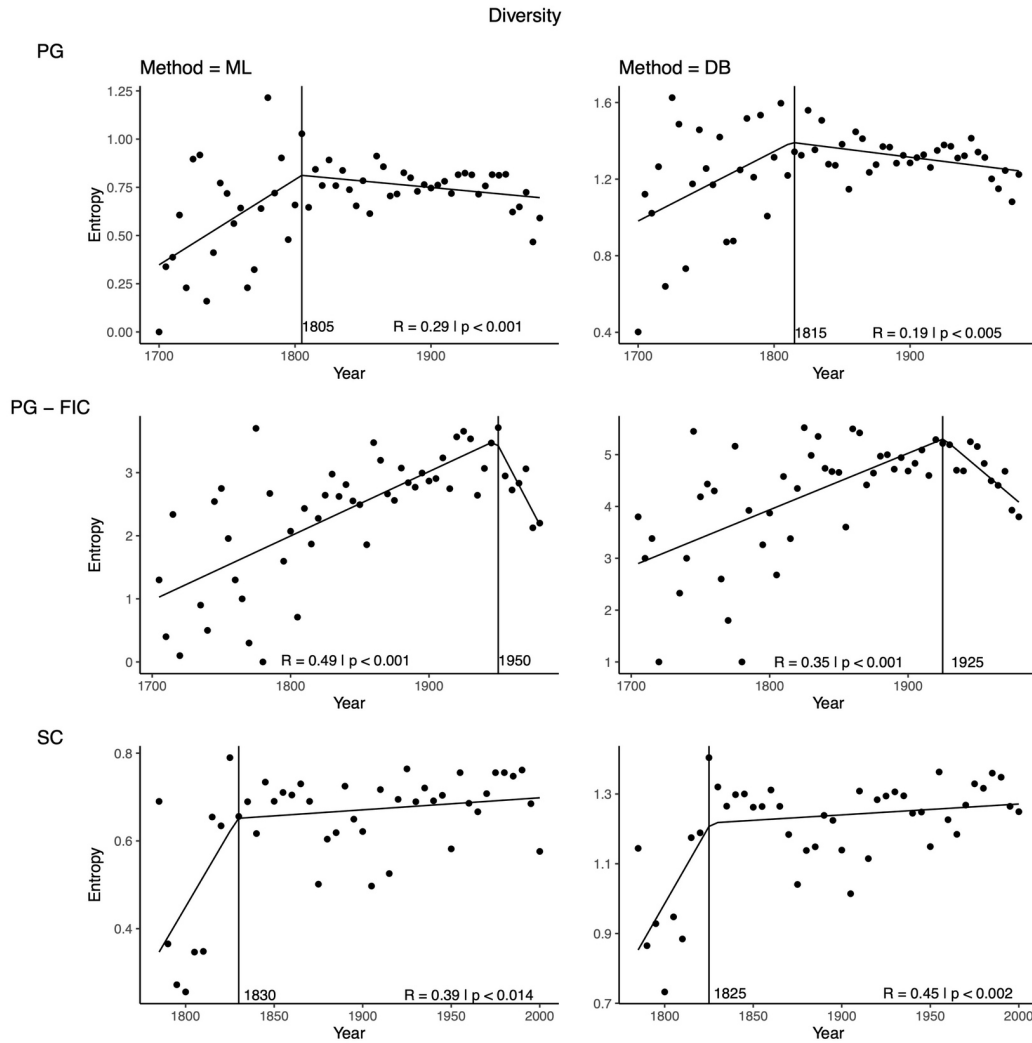


Figure 6. Diversity measures comparing the two methods

due to the inclusion of non-fiction in their data, which behaves considerably differently from a stylistic point of view (cf. Row 4 of [Fig. 4](#)). Why this is the case we leave to future work as well as the further study of other literary genres like poetry or drama. The present work and that of Langer et al. cannot speak to the place of biological taxa in the history of poetry or the theatre, both of which would be valuable further undertakings.

This exercise raises two key points for future work in cultural analytics. The first, which has been well-discussed in the field, is that corpus construction matters considerably. How we define our objects of study impacts greatly what we observe. I would argue that the term “creative literature” is too broad and that focusing on the effects of specific genres and modes of writing – i.e. paying greater attention to the effects of what Genette calls *architextuality* (Genette) – is essential for future work in cultural analytics. As is clear with the data used here, different genres behave differently with respect to the rate of taxa. Speaking of a trend with respect to “creative literature” that is directionally opposite to that exhibited by fiction, the dominant component of this concept, is historically misleading.

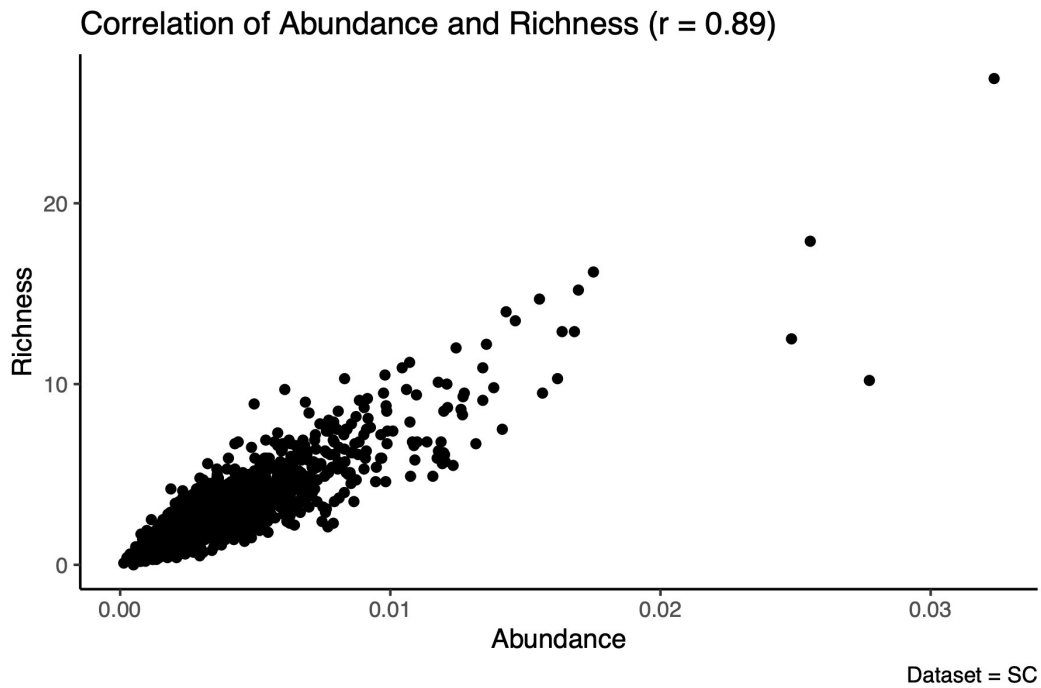


Figure 7. Correlation plot of Abundance and Richness for the SC data using the ML method.

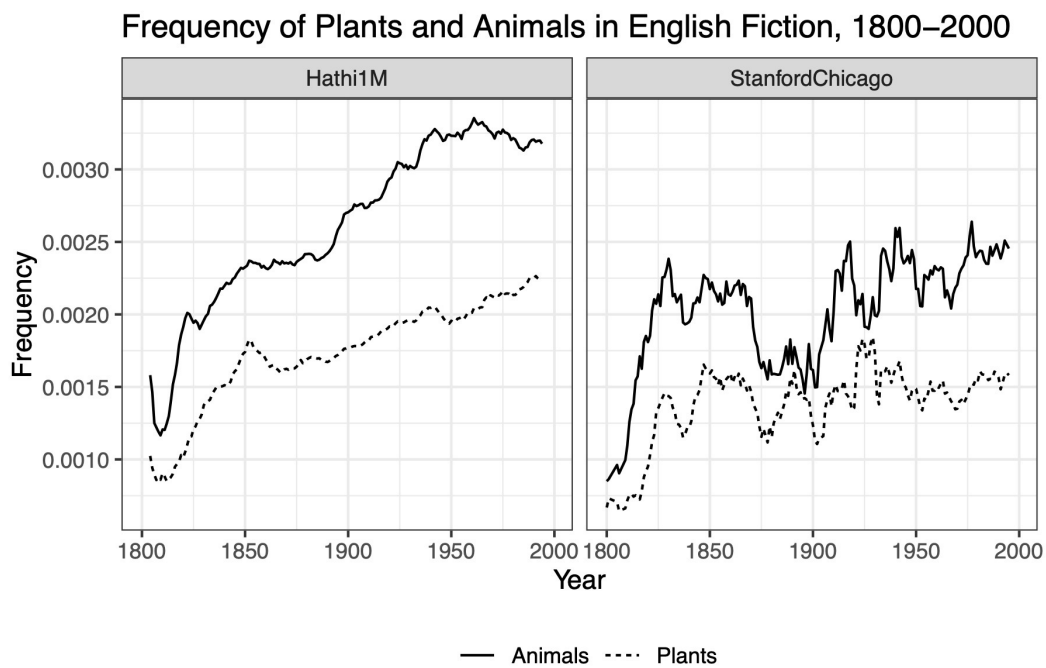


Figure 8. Relative abundance of plants and animals in the different fiction datasets.

At the same time, the current work also affirms the argument made in prior work (Underwood et al.) that using multiple datasets can give us confidence of larger historical trends. While no one sample can perfectly capture the history of such large cultural domains as human writing, when we see multiple samples of writing behave in directionally similar ways we can gain confidence

regarding long-term historical trajectories. Given the data here, it would indeed be surprising if fictional references to biological taxa were ultimately found not to have increased over the past two centuries at least in English.

This finding raises a second more complex matter. As Langer et al. rightly point out, the indication of a long-term semantic dis- / investment in biological entities in the written record would indeed be cause for concern / hope. It would help shore up beliefs in the gradual alienation / attachment of humans from / to the natural world as the stories they tell about themselves include fewer / more remnants of the Earth's biome. Just as Langer et al. argue that their findings confirm widespread beliefs in modern humans' alienation from nature, one could argue that the findings here support the idea of modern humans' *attachment* to nature through a continually growing investment in describing the natural world through imaginative storytelling.

The problem that remains in both cases is whether the rise or fall of biological taxa in any kind of writing is indicative of "social attitudes" towards the natural world. This problem compounds when we are talking about works of fiction where the entities do not strictly refer to the real world.

While the present work cannot address these problems directly, the findings here point to one possible interpretation for future work. The dramatic rise of taxa in the first half of the nineteenth century followed by far more gradual and consistent deployment of taxa over the ensuing decades in English-language fiction may not be capturing a broader investment in "nature" so much as a major stylistic shift that transpires in the first half of the nineteenth century that we could label as "narrative concretization" and that prior research has gestured towards (Heuser and Le-Khac; Underwood, *Distant Horizons*). Rather than explain the rapid growth of biological taxa in one period followed by its relatively steady increase thereafter as an index of a growing social commitment to the natural world (which runs counter to our expectations of modernization and industrialization), we might alternately consider these trends as indices of the social function of fiction as a mode of writing that increasingly foregrounds materially concrete experience in the modern marketplace of ideas.

Nature in fiction would thus not be understood referentially – as a sign of nature – but far more as a "prop," one type of entity among many others that populate fiction and lend it a sense of concretization. A further piece of evidence to suggest this view is the observation made by Langer et al. that taxa references in fiction tend to be highly generic rather than specific and descriptive (see [Figure 2](#)). Rather than invoke taxonomic specificity about animals or plant life, fiction relies more heavily on generic entities such as "horse", "pig", or "tree." Further work should explore the extent to which generic types are reflective of social attitudes or placeholders of other kinds of meaning. The dramatic increase of taxa prior to 1835/1850 may simply be a sign of a reorientation of fictional writing around a particular mode of

representation that achieves long-term consensus due to the emergence of particular reading audiences and market conditions. Biological taxa would thus be indicators of something happening to “fiction” rather than “social beliefs” about nature. We can refer to the model proposed by Langer et al. as the “mimetic theory” of biodiversity – where language functions as an indicator of social attitudes – and the one proposed here as the “formal theory,” where the presence of biological taxa have a function unrelated to their literal meaning, i.e. they do other kinds of work for authors than signal social beliefs about nature. It is worth emphasizing that these theories are not mutually exclusive.

The project by Langer et al. establishes a valuable research program and a set of resources to better understand human-nature implication. The data and methods presented here can help augment this prior work by providing more reliable estimates of the behavior of biological taxa in fiction, offering an important baseline against which any observable shifts in the future occur. It may well be the case that the stylistic consensus we have observed for the past 150 years will undergo a major transformation in the decades to come due to notable climatic shifts. Of one thing we can be certain: the stories we tell about the Earth are going to change.

Data Repository: <https://doi.org/10.7910/DVN/KICQOU>

Peer Reviewers: *Thierry Poibeau & Mohit Iyyer*

Submitted: August 19, 2022 EST, Accepted: September 26, 2022 EST



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at <http://creativecommons.org/licenses/by/4.0> and legal code at <http://creativecommons.org/licenses/by/4.0/legalcode> for more information.

REFERENCES

- Argamon, Shlomo, et al. "Gender, Genre, and Writing Style in Formal Written Texts." *Text & Talk*, vol. 23, no. 3, 2003, pp. 321–46.
- Bagga, Sunyam, and Andrew Piper. "HATHI 1M: Introducing a Million Page Historical Prose Dataset in English from the Hathi Trust." *Journal of Open Humanities Data*, vol. 8, 2022, [doi:10.5334/johd.71](https://doi.org/10.5334/johd.71).
- Bamman, David, et al. "A Bayesian Mixed Effects Model of Literary Character." *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, [doi:10.3115/v1/p14-1035](https://doi.org/10.3115/v1/p14-1035).
- Bode, Katherine. "Why You Can't Model Away Bias." *Modern Language Quarterly*, vol. 81, no. 1, Mar. 2020, pp. 95–124, [doi:10.1215/00267929-7933102](https://doi.org/10.1215/00267929-7933102).
- Genette, Gérard. *The Architext: An Introduction*. Univ of California Press, 1992.
- Heuser, Ryan, and Long Le-Khac. "A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method." *Stanford Literary Lab Pamphlet*, vol. 4, 2012.
- Kesebir, Selin, and Pelin Kesebir. "A Growing Disconnection from Nature Is Evident in Cultural Products." *Perspectives on Psychological Science*, vol. 12, no. 2, Mar. 2017, pp. 258–69, [doi:10.1177/1745691616662473](https://doi.org/10.1177/1745691616662473).
- Langer, Lars, et al. "The Rise and Fall of Biodiversity in Literature: A Comprehensive Quantification of Historical Changes in the Use of Vernacular Labels for Biological Taxa in Western Creative Literature." *People and Nature*, vol. 3, no. 5, Sept. 2021, pp. 1093–109. *Crossref*, [doi:10.1002/pan3.10256](https://doi.org/10.1002/pan3.10256).
- Lee, James Jaehoon, and Joshua Beckelhimer. "Anthropocene and Empire: Discourse Networks of the Human Record." *PMLA*, vol. 135, no. 1, Jan. 2020, pp. 110–29, [doi:10.1632/pmla.2020.135.1.110](https://doi.org/10.1632/pmla.2020.135.1.110).
- Piper, Andrew. *Enumerations*. University of Chicago Press, 2018, [doi:10.7208/chicago/9780226568898.001.0001](https://doi.org/10.7208/chicago/9780226568898.001.0001).
- . "Fictionality." *Journal of Cultural Analytics*, vol. 2, no. 2, Dec. 2016, [doi:10.22148/16.011](https://doi.org/10.22148/16.011).
- Piper, Andrew, and Eva Portelance. "How Cultural Capital Works: Prizewinning Novels, Bestsellers, and the Time of Reading." *Post45*, vol. 10, 2016.
- Sap, Maarten, et al. "Recollection versus Imagination: Exploring Human Memory and Cognition via Neural Language Models." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, [doi:10.18653/v1/2020.acl-main.178](https://doi.org/10.18653/v1/2020.acl-main.178).
- Schmidt, Benjamin, et al. "Uncontrolled Corpus Composition Drives an Apparent Surge in Cognitive Distortions." *Proceedings of the National Academy of Sciences*, vol. 118, no. 45, Nov. 2021, p. 2115010118, [doi:10.1073/pnas.2115010118](https://doi.org/10.1073/pnas.2115010118).
- Travis, Charles. "The Digital Anthropocene, Deep Mapping, and Environmental Humanities' Big Data." *Resilience: A Journal of the Environmental Humanities*, vol. 5, no. 2, 2018, p. 172, [doi:10.5250/resilience.5.2.0172](https://doi.org/10.5250/resilience.5.2.0172).
- Underwood, Ted. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press, 2019, [doi:10.7208/chicago/9780226612973.001.0001](https://doi.org/10.7208/chicago/9780226612973.001.0001).
- . "Genre Theory and Historicism." *Journal of Cultural Analytics*, vol. 2, no. 2, Oct. 2016, [doi:10.22148/16.008](https://doi.org/10.22148/16.008).
- . "NovelTM Datasets for English-Language Fiction, 1700-2009." *Journal of Cultural Analytics*, vol. 5, no. 2, May 2020, p. 13147, [doi:10.22148/001c.13147](https://doi.org/10.22148/001c.13147).

Wilkins, Matthew. "Genre, Computation, and the Varieties of Twentieth-Century U.S. Fiction."
Journal of Cultural Analytics, vol. 2, no. 2, Nov. 2016, [doi:10.22148/16.009](https://doi.org/10.22148/16.009).