Received: 1 February 2021 | Accepted: 24 June 2021

DOI: 10.1002/pan3.10256

PEOPLE AND NATURE BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

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

Lars Langer[1,2] | Manuel Burghardt[3] | Roland Borgards[4] | Katrin Böhning-Gaese[5,6,7] | Ralf Seppelt[2,7,8] | Christian Wirth[1,7,9]

[1]Department of Systematic Botany and Functional Biodiversity, Leipzig University, Leipzig, Germany; [2]Department Computational Landscape Ecology, Helmholtz-Centre for Environmental Research GmbH (UFZ), Leipzig, Germany; [3]Computational Humanities Group, Leipzig University, Leipzig, Germany; [4]Department of German Literature, Goethe University, Frankfurt am Main, Germany; [5]Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany; [6]Institute of Ecology, Diversity and Evolution, Goethe University, Frankfurt am Main, Germany; [7]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany; [8]Institute of Geoscience & Geography, Martin-Luther-University Halle-Wittenberg, Halle (Saale), Germany and [9]Max-Planck Institute for Biogeochemistry, Jena, Germany

Correspondence
Lars Langer
Email: lars.langer@uni-leipzig.de

Funding information
Universität Leipzig; Helmholtz-Zentrum für Umweltforschung

Handling Editor: Shonil Bhagwat

Abstract

1. Nature's non-material contributions to people are difficult to quantify and one aspect in particular, nature's contributions to communication (NCC), has so far been neglected. Recent advances in automated language processing tools enable us to quantify diversity patterns underlying the distribution of plant and animal taxon labels in creative literature, which we term BiL (biodiversity in literature). We assume BiL to provide a proxy for people's openness to nature's non-material contributions enhancing our understanding of NCC.
2. We assembled a comprehensive list of 240,000 English biological taxon labels. We pre-processed and searched a subcorpus of digitised literature on Project Gutenberg for these labels. We quantified changes in biodiversity indices commonly used in ecological studies for 16,000 books, encompassing 4,000 authors, as proxies for BiL between 1705 and 1969.
3. We observed hump-shape patterns for taxon label richness, abundance and Shannon diversity indicating a peak of BiL in the middle of the 19th century. This is also true for the ratio of biological to general lexical richness. The variation in label use between different sections within books, quantified as β-diversity, declined until the 1830s and recovered little, indicating a less specialised use of taxon labels over time.
4. This pattern corroborates our hypothesis that before the onset of industrialisation BiL may have increased, reflecting several concomitant influences such as the general broadening of literary content, improved education and possibly an intensified awareness of the starting loss of biodiversity during the period of

romanticism. Given that these positive trends continued and that we do not find support for alternative processes reducing BiL, such as language streamlining, we suggest that this pronounced trend reversal and subsequent decline of BiL over more than 100 years may be the consequence of humans' increasing alienation from nature owing to major societal changes in the wake of industrialisation.

5. We conclude that our computational approach of analysing literary communication using biodiversity indices has a high potential for understanding aspects of non-material contributions of biodiversity to people. Our approach can be applied to other corpora and would benefit from additional metadata on taxa, works and authors.

KEYWORDS

biodiversity in literature, computational literary studies, cultural ecosystem services, environmental humanities, historical biodiversity, Nature's Contributions to People, non-material contribution, text mining

1 | INTRODUCTION

Our planet is losing biodiversity at unprecedented rates due to land-use change, direct exploitation, climate change, pollution and the invasion of exotic species (Cardinale et al., 2012; IPBES, 2019; Millennium Ecosystem Assessment, 2005; Tilman, 1999). In Western countries, for example, this loss had started already in the second half of the 18th century, with the onset of industrialisation and modern agriculture (Krausmann & Haberl, 2002; Lambin & Geist, 2006; Ulloa-Torrealba et al., 2020). Ecosystems and their biodiversity contribute to human wellbeing and the functioning of societal subsystems (IPBES, 2019; Millennium Ecosystem Assessment, 2005) in various ways. There is conclusive evidence that biodiversity loss is harmful for ecosystem functioning (Cardinale et al., 2012; Schmid et al., 2009) and consequently for nature's contributions to people (NCP; Díaz et al., 2015, 2018), as the anthropological side of the entangled nature–culture (Barad, 2007; Haraway, 2016; Haraway & Begelke, 2003). Understanding and evaluating NCP is critical to facilitating governments' decision making if they are to achieve a healthy and sustainable future (Díaz et al., 2015; Pascual et al., 2017).

However, assessing the various forms that NCP take is challenging, because the majority of contributions are difficult to quantify (Daily, 2000; Daniel et al., 2012; Martinez-Alier, 2002). Substantial progress has been made to assess material contributions using new data-acquisition tools and modelling approaches (de Araujo Barbosa et al., 2015; Maes et al., 2015) and to quantify their monetary values (Lautenbach et al., 2012; Schmidt et al., 2016; Sumarga et al., 2015). However, non-material contributions—including recreation and education, as well as cultural and religious uses, but also nature's aesthetic value and the appreciation of acknowledging nature as a necessary complement to human culture—remain difficult to measure and cannot be comprehensively and reliably reduced to a similar single value (Lautenbach et al., 2019; Pascual et al., 2017; Seppelt et al., 2011).

Fundamental to understanding non-material contributions, especially towards human wellbeing, is quantifying people's valuation of biodiversity. Recent attempts include surveys (Ainscough et al., 2019), qualitative and quantitative assessments of the disconnection from nature in cultural products (Celis-Diez et al., 2016; Kesebir & Kesebir, 2017; Prévot-Julliard et al., 2015) and analyses of knowledge loss within graphical media (Wolff et al., 1999). The recently established field of conservation culturomics further enhances our understanding by analysing word frequencies of limited word lists within mainly contemporary corpora (e.g. from social media platforms). It does so to approximate ongoing cultural change (Michel et al., 2011) as a means of comprehending and eventually predicting public interest in certain species, areas and human–nature interactions (Ladle et al., 2016; Willemen et al., 2015).

Despite these efforts, there is still a fundamental lack of understanding regarding the influence of nature on various aspects of our culture and its development. Culture can be seen as interpersonally transferred information (Mesoudi, 2011) that requires communication to exist and function, a subject which so far has received little attention in research on NCP. Our communication is certainly influenced by the appearance and behaviour of animals and plants and by their semiotics (Tüür & Tønnessen, 2014), as in the floriography (language of flowers) of the 19th century, where plant taxa were related to personal character traits (Gagliano et al., 2017; Greenaway, 1884), or in the naming of precise colours after certain plants or animals, as with 'violet' or 'vermilion' (from Latin for 'worm') respectively. This biodiversity-aided precision and sophistication of communication may be critical in the development of a socially and technically advanced civilisation, and thus facilitates a high standard of human wellbeing.

To address this critical lack, we present an approach that extracts and analyses biodiversity information from creative literature, which here is defined as the category combining works of fiction (constituting the preponderance of our corpus), travelogues, biographies and letters. We recognise literature as one important form of long-preserved communication that allows us to quantify the degree to which our culture uses labels for biological taxa within communication over time. 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. We recognise this as a part of nature's contribution to communication and as a component of non-material contributions by nature to people.

In order to quantify the usage of taxon labels and investigate its diachronic development, we analysed a corpus of 16,000 digitised literary works covering nearly 300 years of Western literature in English, including English translations of works originally published in other languages. This corpus contains more than 1.2 billion tokens (words) that were searched for about 240,000 English taxon labels. In contrast to related studies (Kesebir & Kesebir, 2017; McCrindle & Odendaal, 1994; Queiroz et al., 2015; Willemen et al., 2015; Wolff et al., 1999) that either only analysed a small fraction of the size of our corpus or searched only for a limited set of taxon labels (an order between 1 and 10^2 labels), we aimed for a comprehensive investigation that tried to find every non-human living being that is mentioned in the corpus. This comprehensiveness was the precondition for calculating the diversity of taxon labels using both richness (number of taxon labels) and diversity indices, for example Shannon diversity (Magurran & McGill, 2011, Chapter 5), and for evaluating this biodiversity in literature (BiL) against the background of changes in general lexical richness, which we determine by the number of types (unique tokens). It also allowed us to calculate several facets of biodiversity borrowing from ecological theory (Magurran & McGill, 2011, Chapter 7). By distinguishing diversity at different scales (local α-diversity at the 'plot scale', i.e. a book section of 1,000 words, and γ-diversity of 'regions', i.e. a size-normalised work), we were able to calculate the sequential dissimilarity (β-diversity), that is, the dissimilarity of taxon labels used in different sections of a book. Consequently, we can quantify how pervasive or, in contrast, exceptional the use of taxon labels is.

We acknowledge that biodiversity and its distribution in texts is likely to be controlled by a complex set of drivers (e.g. the percentage of authors socialised in urban environments, changing poetological/narrotological norms in different times and different cultures and historical changes in the social function of creative literature) which we currently cannot approximate quantitatively. For this reason, we have deliberately avoided disentangling causal relationships at this stage. Furthermore, BiL does not only reflect authors' biodiversity awareness, but may also be driven by general processes of language transformation, such as streamlining of vocabulary. Alongside the description of temporal trends in components of BiL, we nevertheless put forward a hypothetical scenario to discuss our findings in a qualitative fashion: We expect that the growing disconnection from nature, induced by industrialisation, urbanisation and extensive land-use change at the onset of industrial agriculture and forestry in the 19th century (Brown & Harrison, 1978, Chapter 2; Grigg, 1987; Seppelt & Cumming, 2016), is temporally correlated with a decrease in BiL towards the end of the 19th century. With our time series starting in the early 1700s, we also hypothesise that the usage of taxon labels initially increases during the time of enlightenment, which promoted the natural sciences and the educational system, and romanticism, which has partly been interpreted as a proto-ecological countermovement opposing the industrialisation of life (Trepl, 1987), and which begins to understand nature as a complex system of interrelated and interdependent dynamic elements (Detering, 2020, pp. 307–370; Morton, 2007; Rigby, 2014, 2020), reaching a peak in the 1830s.

2 | METHODS

The process for obtaining the necessary data involves several steps, which we illustrate in Figure 1 and detail in the following sections.

2.1 | Corpus

A corpus reflecting cultural dynamics (Ladle et al., 2016) has to contain a sufficiently extensive proportion of literature to enable us to draw random subsamples throughout the time span analysed, in our case, 1705–1969. We chose this period for three reasons: (a) It provides sufficient digitised print products for building a corpus; (b) its starting point predates the peak of the industrial and agricultural revolution and thus allows us to study its potential effect over time; (c) its endpoint predates the digital revolution that has fundamentally changed access to knowledge with potential consequences for label usage in literature. Subsamples from a corpus have to be large enough to represent the population of texts of a given time period, thereby avoiding biases introduced by idiosyncrasies of authors' attitudes, regions, genres and text types. In most available corpora, creative literature is underrepresented or biased towards only a few canonical authors or works.

For its extensive and random (i.e. non-selective) content, as well as its open access, which benefits the reproducibility of our project, we are working with the corpus of Project Gutenberg. We downloaded the Standardized Project Gutenberg Corpus (SPGC; Gerlach & Font-Clos, 2020) in February 2019. The complete downloaded SPGC corpus contained approximately 59,000 works. The SPGC contains a file for basic metadata provided by uploaders to the Project Gutenberg Library.

However, these metadata do not contain the works' years of publication. Therefore, we applied an estimation by using the dates of birth and death of the corresponding authors as given by Project Gutenberg. The formula applied was: $\text{year} = 0.5 \times (\text{birth} + 21 + \text{death})$, corresponding to the central year between the starting age of 21 and the year of death. The parameters of the formula were derived by screening sensible values for a starting age and ratios between this starting age and death. For each combination, we determined the standard deviation from the actual year of publication, based on randomly gathered metadata for a subset of 4,705 works. We attained the lowest standard deviation of approximately 6.9 years with the parameters of a starting age of 21 and the ratio of 0.5.

From the 59,000 works, we selected all works with an estimated publication date between 1705 and 1969. We excluded all items that were not categorised as English text, thereby also including all

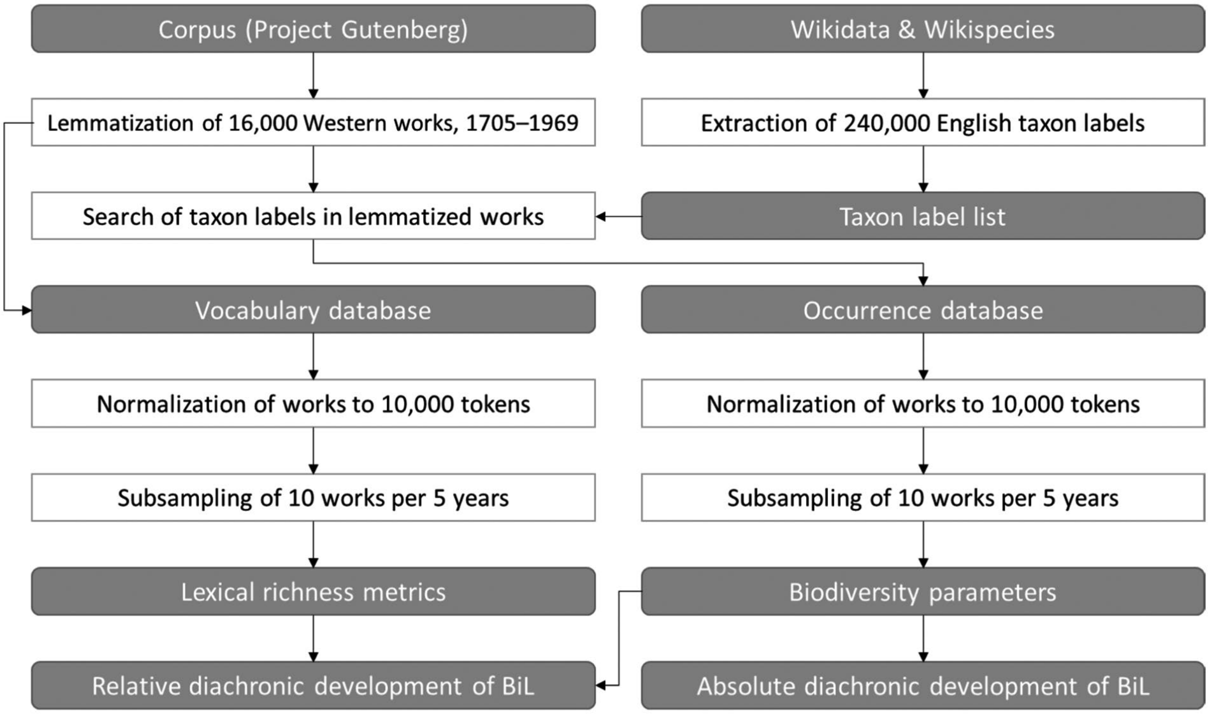

Corpus (Project Gutenberg)
Lemmatization of 16,000 Western works, 1705–1969
Search of taxon labels in lemmatized works
Vocabulary database
Normalization of works to 10,000 tokens
Subsampling of 10 works per 5 years
Lexical richness metrics
Relative diachronic development of BiL
Wikidata & Wikispecies
Extraction of 240,000 English taxon labels
Taxon label list
Occurrence database
Normalization of works to 10,000 tokens
Subsampling of 10 works per 5 years
Biodiversity parameters
Absolute diachronic development of BiL

FIGURE 1 The key steps in estimating an absolute and relative trend of biodiversity in literature (BiL) from the selection of 16,000 Western works of the Project Gutenberg corpus and the databases of Wikidata and Wikispecies as sources for taxon labels

English translations from other languages. The subject flags within the metadata were searched for specific keywords (e.g. 'fiction', 'novel' and 'travel', see Appendix A1) in order to include only authors that published creative literature. We included only works that were unambiguously ascribed to a single author, as opposed to institutions (e.g. departments, universities and journals, see Appendix A2). For each text, we determined its length by the number of individual tokens and its vocabulary size by the number of unique tokens. As we applied subsampling for bias reduction, we chose to only include works with a length of 15,000 words or more to allow for a representative size of our subsamples as described in Section 2.4. The final corpus contains 15,798 works by 3,832 authors.

Plotting these works shows that the SPGC contains an uneven temporal distribution, as shown in Figure 2. This distribution mainly reflects the status of digitisation of books generally and the initially low publishing rates of fiction prior to the 19th century, and corresponds to the strong increase in population in the regions related to the corpus, mainly Europe and North America, throughout the investigated time period (Biraben, 2003). Additionally, the first half of the investigated period sees the successive development of a prominent middle class as the main producers for literature (Hudson, 2015). The decline within the 20th century mainly reflects the limited availability of works within the public domain, which in turn depends on copyright expiration dates. In the United States, the Copyright Term Extension Act of 1998 determines that a work published within the investigated period typically enters the public domain 95 years after publication. In specific cases and countries, this date may also depend on the death of the author.

Further preparation of our corpus was required as the label database contains only the non-inflected base form (lemma) for each taxon label (see Section 2.2). To overcome this we used the Lemmatiser of the Stanford CoreNLP Toolkit (Manning et al., 2014) to normalise all tokens within the corpus to their lemma. This step allowed the search algorithm to find inflected taxon labels.

2.2 | Label database

Our label database aims to contain the largest possible number of English biological taxon labels to allow us to calculate meaningful diversity indexes (see below). The bases for the label database are the open encyclopaedias Wikispecies and Wikidata (Vrandečić & Krötzsch, 2014), downloaded in February 2019. From both sources, we collected all taxonomic data. Subsequently, we extracted and compiled all English labels, which we here call taxon labels. However, because we could not rule out that the authors represented within the corpus may have used spelling variants of taxon labels, in addition to the taxon label at hand, we included all spelling variants that would result from replacing spaces by hyphens and vice versa as well as from omitting spaces, apostrophes, hyphens and leading adjectives altogether. We reduced the character set of all taxon labels to the English alphabet (e.g. by simplifying letters with diacritics), retaining only spaces, apostrophes and hyphens as additional symbols.

Furthermore, we manually produced a blacklist (Appendix A3) containing ambiguous labels with a probability of generating false positives due to homography (e.g. 'bishop', 'diver' and 'ray') or

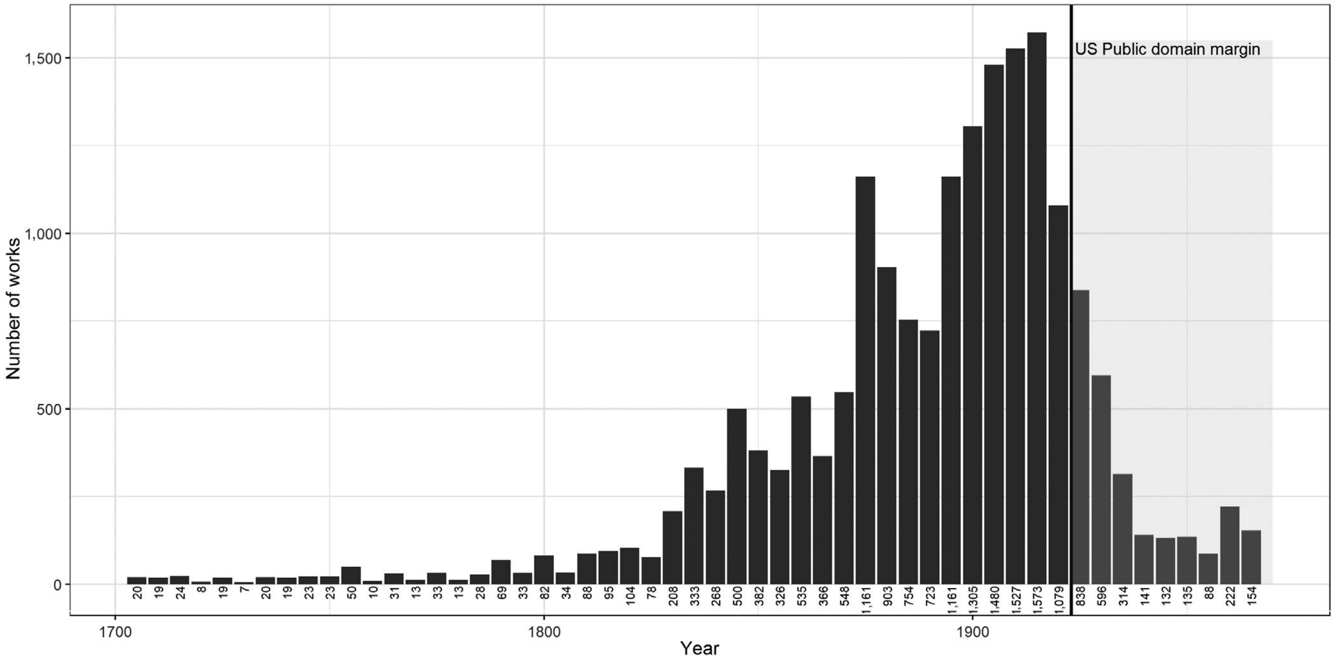


FIGURE 2 Temporal distribution of works. The metadata of the corpus did not contain the year of publication, but the authors' biographical data. Hence, we estimated the publication year and placed the works into intervals of 5 years. For reference, we marked the typical US margin for works entering the public domain, which significantly contributes to the decrease in the number of works within this openly available corpus

artefacts in the label extraction (e.g. 'european', 'alexander' and 'red'), as a result of indistinguishable formatting in the original data. In order to match a label from the label list with a token within the corpus the sequences of the characters had to be identical. This is why automated text extraction, in particular when concerning unusual tokens such as rarely used species names, is generally dependent on the accuracy of the individual uploaders when formatting (e.g. consistent word order, use of parentheses and correct orthography) and providing the information via openly available encyclopaedias. During the extraction of the taxon label, we produced several false findings, like adjectives, locations or names, which we also included in the blacklist. The resulting database contains entries throughout the biological domains and taxonomic levels.

After following the above steps, our label database contains 161,488 entries extracted from Wikispecies, to which we added another 80,955 entries extracted from Wikidata, totalling 242,443 taxon labels. 35,588 of these labels are synonyms and a further 106,592 labels are spelling variants. Altogether, they refer to 100,263 taxa; 214,941 of the taxon labels refer to 91,244 taxa at the species level.

2.3 | Search

In order to get a comprehensive quantification of BiL, we searched our corpus for all taxon labels. For consistency with the label database, we reduced the character set of the corpus to the English alphabet in the same manner as with the label database. We preserved the following additional characters: spaces, new lines, apostrophes, hyphens, full stops, exclamation marks, question marks, commas, semicolons, quotation marks and numbers.

Each of the 1,559,771 frames within all works was searched for all the taxon labels from our database. When a label was found within the text, an occurrence entry was generated comprising the label, scientific name, work and frame number. The result was a database comprising all occurrences of BiL.

2.4 | Analytical methods

For an analysis of the diachronic development of BiL, we grouped the works into intervals of 5 years, according to their estimated year of publication, starting with the year 1705. For each interval, we randomly selected 10 works, each by a different author. If the corpus contained fewer than 10 authors in a 5-year interval, we selected all of them and randomly selected one of their works. As this process is prone to chance and may result in an extreme biodiversity measurement, we repeated this process one hundred times for each interval and averaged the determined parameters, which we introduce below.

A frame corresponds to a plot of some predetermined area in typical ecological investigations (Magurran & McGill, 2011, Chapter 2). Therefore, we regard a frame as an α-region where α-diversity can be determined. As several biodiversity measures are influenced by the length of works, similar to a species-area effect in ecology, we applied a normalisation by randomly selecting 10 such frames

TABLE 1 Biodiversity parameters determined for a normalised work (i.e. per 10 randomly sampled frames), as resulting from the arithmetic mean of 100 sampling iterations. The frame corresponds to an α-region, a plot in real-life experiments. The normalised work corresponds to a γ-region

Biodiversity parameter	Quantified for	Description
Richness	1,000-word-frame Normalised work	Number of unique taxon labels (types)
Abundance	Normalised work	Total number of all taxon labels, including repeated mentions (tokens)
Shannon Diversity according to the Shannon–Weaver Index	1,000-word-frame (α-diversity) Normalised work (γ-diversity)	$H = -\sum_{i=1}^{S} p_i \ln p_i$ With iteration through all unique taxon labels (1 → S) and p the proportion of that label to all taxon labels in the respective section
Beta diversity	Normalised work	$\beta = \frac{\gamma}{\mathrm{mean}(\alpha)}$ A measure of dissimilarity, based on Whittaker (1960), between frames within one normalised work

for each work. Again, the process was iterated one hundred times for each work to be able to average the determined parameters. A complete normalised work is referred to as the γ-region and is characterised by the γ-diversity.

For each work we determined the abundance of taxon labels, both richness and Shannon diversity for cross-comparison, and β-diversity as a measure of the dissimilarity of taxon labels between frames within works. The exact parameters determined for each work are given in Table 1. To avoid bias in this process and to leave space for a random selection, we excluded all works with fewer than 15 frames (corresponding to 15,000 words).

Language streamlining could reduce BiL by decreasing the variety of synonyms for taxon labels. To obtain a proxy for this process we compared BiL trends based on the richness of taxon labels (as above) with those based on the overall richness of terms related to distinguished biological taxa without separating synonyms for the same taxon, within 5-year periods. If language streamlining is strong in a particular period, we should expect to see a decrease in the ratio of all taxon labels (counting synonyms separately) to distinguished biological taxa. We chose to measure this taxon richness within the whole 5-year period to include the effect of synonymy among different works from the respective period. To avoid size-related biases we randomly sampled 100 frames, corresponding to 10 works with 10 frames each, in the other measurements for each period repeatedly one hundred times.

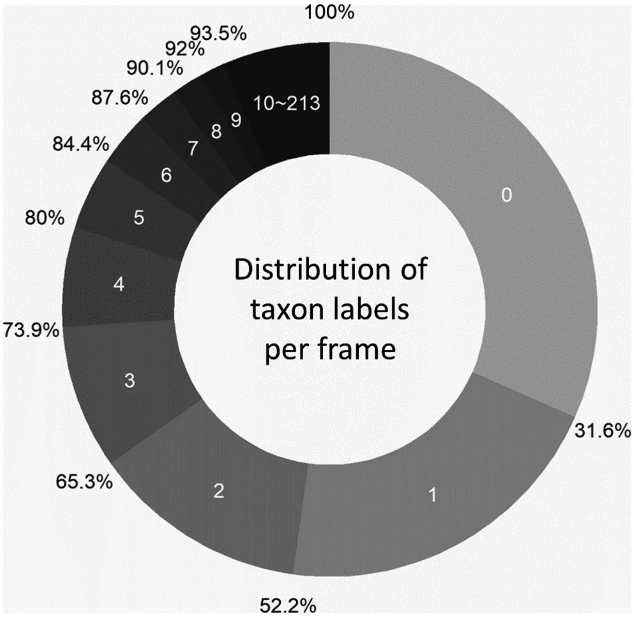


FIGURE 3 Number of frames within the complete works of the corpus that mention a specific number of taxon labels (0 to 9 and ≥10). 31.6% of all frames contained no taxon label and 6.5% of all frames contained 10 or more taxon labels. The maximum is 213 taxon labels in one frame contained in a chapter on animals in a book of jokes (Cole, 1879). Only one other frame within the corpus, contained in a book of animal anecdotes (Goodrich, 1845), exceeds 100 taxon labels

3 | RESULTS

3.1 | Occurrences of taxon labels

More than two thirds of all analysed frames contained at least one taxon label (1,066,839 of 1,559,771 frames; Figure 3). In total, the search revealed 4,416,187 occurrences of 5,994 different taxon labels that refer to 4,652 distinguished biological taxa. One third of these occurrences (1,778,885 of 4,416,187) refer to two thirds of all identified taxa on the species level (3,076 of 4,652). This shows a significant difference between taxon labels at the species level used (one third of occurrences) and species evidently known to authors (two thirds of taxa). Instead of being referred to on the folk-biological species level (Atran & Medin, 2008), most specific organisms are presented on the folk-biological generic level, roughly equating to the scientific genus or family level, as with 'oak', 'eagle' or 'deer', but typically referring to a certain locally dominant species, as with 'European oak', 'golden eagle' or 'red deer' respectively.

Similar to a rank abundance curve in ecology, Figure 4 shows the one hundred most abundant taxon labels in the corpus, representing about two thirds of all occurrences of taxon labels. There is a strong tendency towards labels on either the folk-biological life-form level (Atran et al., 1997) or higher, as with 'tree', 'bird' and 'animal', or the generic level for living beings that humans frequently

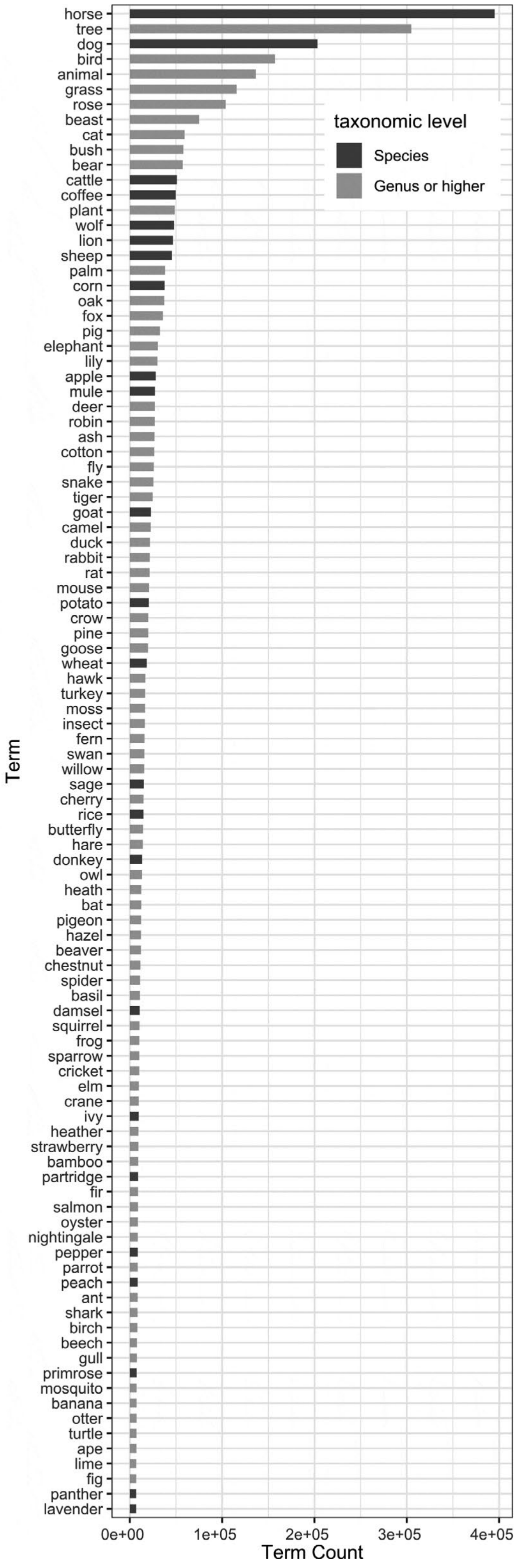


FIGURE 4 The one hundred most abundant taxon labels within the SPGC, totalling 2,872,120 occurrences. Whether a label denotes a taxon at species level or higher is marked by light or dark grey bars, respectively

interact with, especially domesticated animals and plants, like 'horse', 'rose' or 'coffee', and those that pose threats, like 'bear', 'wolf' and 'lion'.

As an analogue to species-area relationships, we expected to find a saturation of label richness with the length of works. By fitting a Michaelis–Menten curve, we determined the half-saturation constant at about 139,000 tokens and a full saturation of the label richness to be at approximately 190 unique taxon labels (Figure 5). Although the plot appears to be scattered, the fit of the Michaelis–Menten function is highly significant ($p < 0.001$).

We also found that authors differ substantially in their use of taxon labels within their œuvre. Generally, authors with an exceptionally large vocabulary (40,000+) do not have a taxon label content of less than 1%. The author with the highest biodiversity vocabulary within our corpus is the 19th-century English novelist Charlotte Mary Yonge with 903 taxon labels. Figure 6 illustrates the relationship between the biodiversity vocabulary and their total vocabulary for each author within the corpus, indicating that the mean percentage of biodiversity vocabulary is 1.09 ± 0.46% of the total vocabulary across authors. Variation across authors is high, ranging from 0.37% to 2.22% (percentiles of 2.5 and 97.5 respectively).

3.2 | Biodiversity parameters

To understand the temporal trends in the usage of taxon labels in works of creative writing in the period under study, we determined size-normalised estimates of richness, abundance, β-diversity for dissimilarity within works and γ-diversity of works. As we observed a clear peak in the 1830s for all our measures, we fitted a change point regression model into each graph.

After a period of increase the development of richness of taxon labels reaches a change point in 1835 and then decreases (Figure 7, top left; $R^2 = 0.59$, $p < 0.001$). We identified a slightly different trend for the development of the abundance of taxon labels (Figure 7, top right): after reaching a maximum in 1836, the average abundance remains almost constant until around 1955 before it decreases abruptly, resulting in a slight overall decrease during the second part of the investigated period ($R^2 = 0.58$, $p < 0.001$). Combining this information with the results for richness, this shows that the authors between 1835 and 1955 on average used a similar number of taxon labels in their works (abundance), but they did not maintain the same level of different taxon labels (richness), which induces an increasing redundancy of taxon labels after 1835.

The development of γ-diversity (Figure 7, upper centre left), quantified as Shannon diversity, falls in line with and supports the results for richness. After the period of increase, the development reached a change point in 1837 and then decreased ($R^2 = 0.58$, $p < 0.001$). By using the γ-diversity, we further determined β-diversity (Figure 7, upper centre right) according to Whittaker (1960) by dividing the γ-diversity per normalised work by the mean of the α-diversity. Here, we observed an inverse development in comparison to the γ-diversity. After an initial decrease the development reaches a change point in 1831 and then slightly increases ($R^2 = 0.55$, $p < 0.001$). At around 1835, γ-diversity is not only at a maximum, but with a minimum in β-diversity the taxon labels are on average more

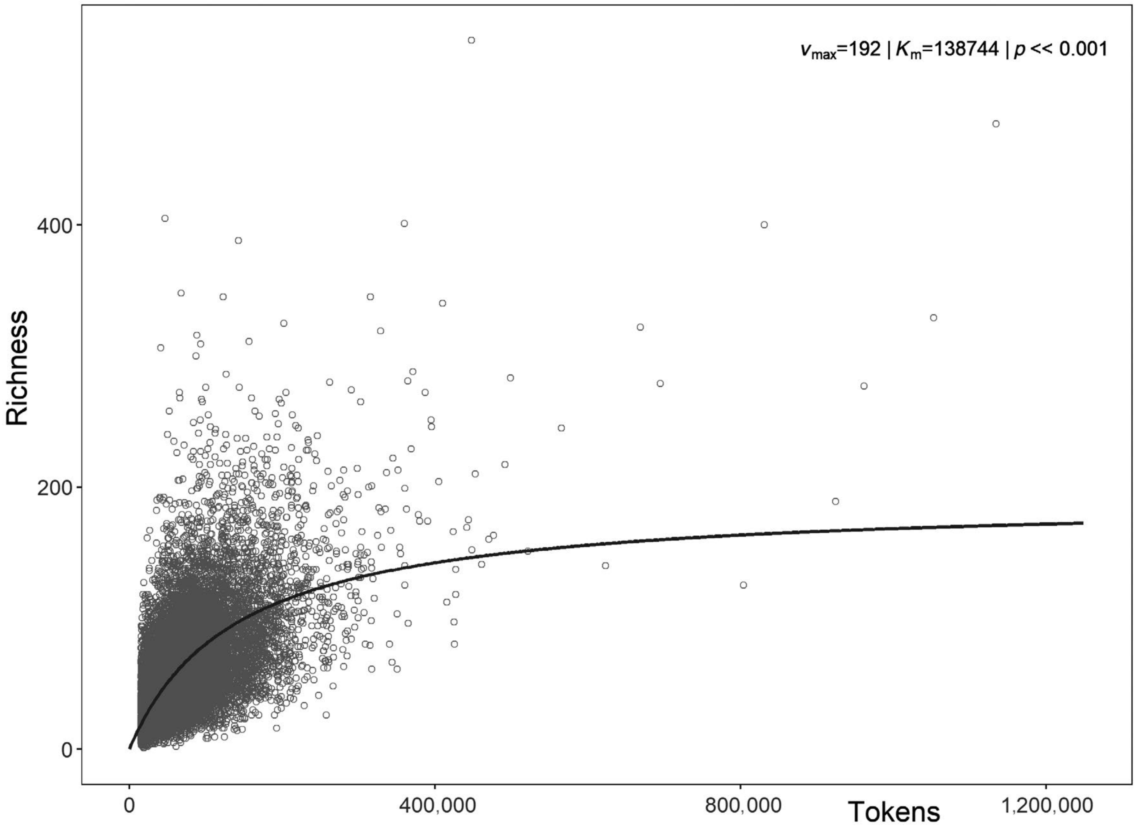


FIGURE 5 Richness of taxon labels throughout each complete work against the length of that work as counted by tokens, showing the similarity to a species-area curve. We fitted a Michaelis Menten function in order to determine the half-saturation constant K_m and the saturation v_{max} of the relationship

uniformly distributed and thus more similar sets of living things are used throughout one work. In the period after 1,830 and until 1950, β-diversity increases only slightly in comparison to the decrease in γ-diversity, meaning that authors on average almost preserved the former degree of similarity of taxon labels in the course of each work, while the overall diversity dwindled, as indicated by the synchronous decline in γ-diversity.

The taxon label richness analysed so far (Figure 7, top left) represents the richness of taxon labels, ignoring the correlation to their actual taxon and thus including and distinguishing synonyms. In Figure 7, lower centre left, we show the richness of distinct biological taxa, the trend of which shows a significant increase in the number of taxa within the first century, followed by a decrease starting around 1805, approximately three decades prior to other similar parameters ($R^2 = 0.46$, $p < 0.001$). In comparison, the richness of taxon labels underwent a slightly steeper decrease but started later. In order to investigate a possible conflation of synonyms over time as a consequence of language streamlining, we calculated the average number of synonyms per taxon and show the corresponding development in Figure 7, lower centre right. The change point analysis in this case yields insignificant results ($R^2 = 0.16$, $p = 0.31$), implying a largely stable count of synonyms per taxon over time.

If we compare the richness trend of taxon labels to the development of the general lexical richness (Figure 7, bottom left) by calculating the ratio of taxon labels' richness and lexical richness (Figure 7, bottom right), we observe a relative decrease in taxon labels. Similar to the development of taxon label richness, the lexical richness increased in the first half of the overall period, reached a change point in 1832 and decreased afterwards ($R^2 = 0.63$, $p < 0.01$). We identified a slightly different trend for the development of the ratio of biological and lexical richness: after reaching a maximum in 1835 following a period of increase, the average abundance only slightly decreased until around 1955 before it decreased abruptly, resulting in an overall decrease during the second part of the investigated period ($R^2 = 0.57$, $p < 0.001$). This development shows that the decrease in taxon labels in the second half of the overall period was stronger than the decrease in lexical richness.

4 | DISCUSSION

By combining a corpus of nearly 16,000 works extracted from the most comprehensive openly available literature corpus with a comprehensive list of labels for living beings, we were able to conduct an analysis of the development of the use of taxon labels within Western creative literature in their corresponding English versions. We show that richness, abundance and Shannon diversity peak in the 1830s, followed by a consistent decline over more than 100 years until the

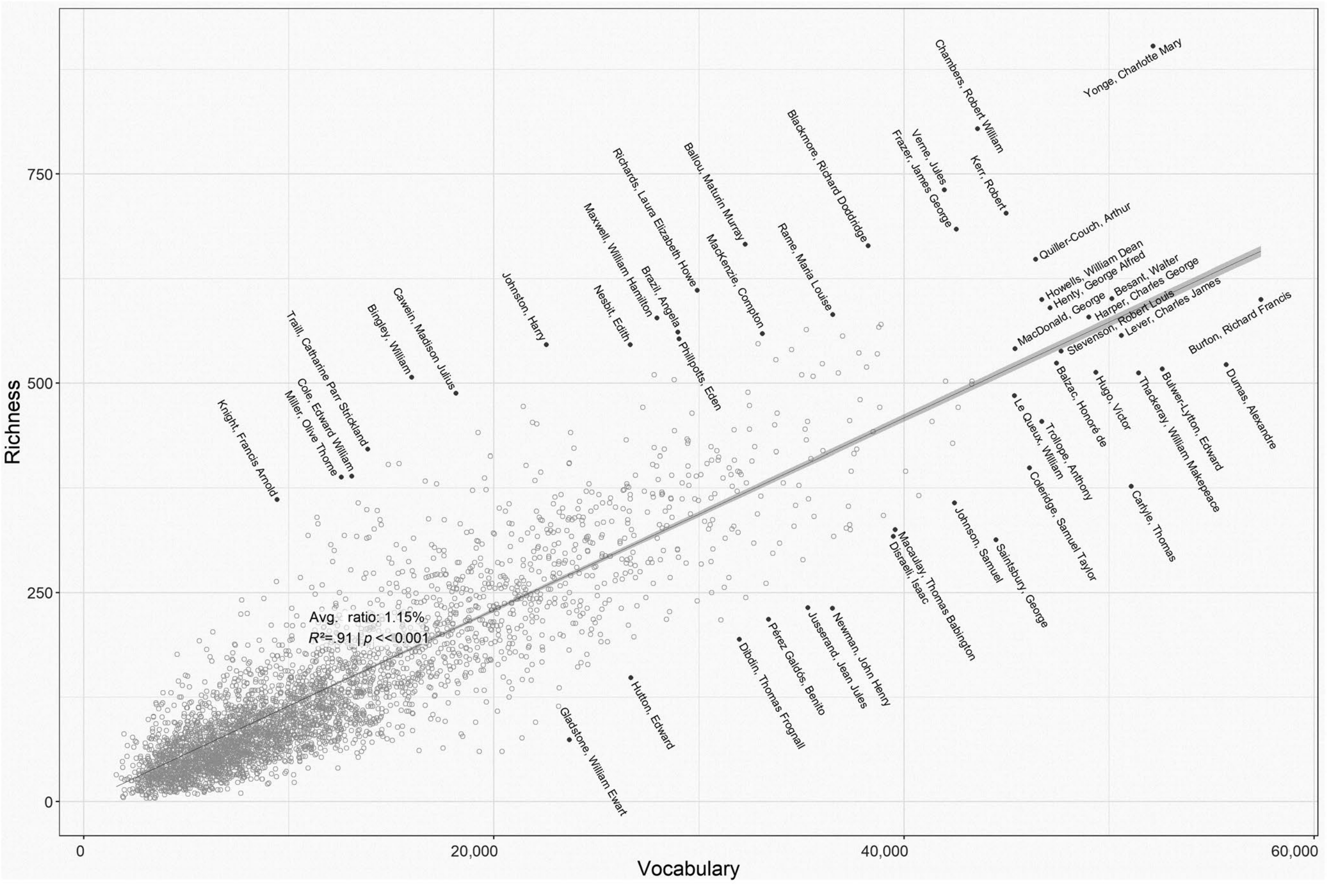


FIGURE 6 Comparison of biodiversity vocabulary with complete vocabulary for each author based on their œuvre as present in our corpus. Individual authors are represented by a dot, outliers and extremes are named

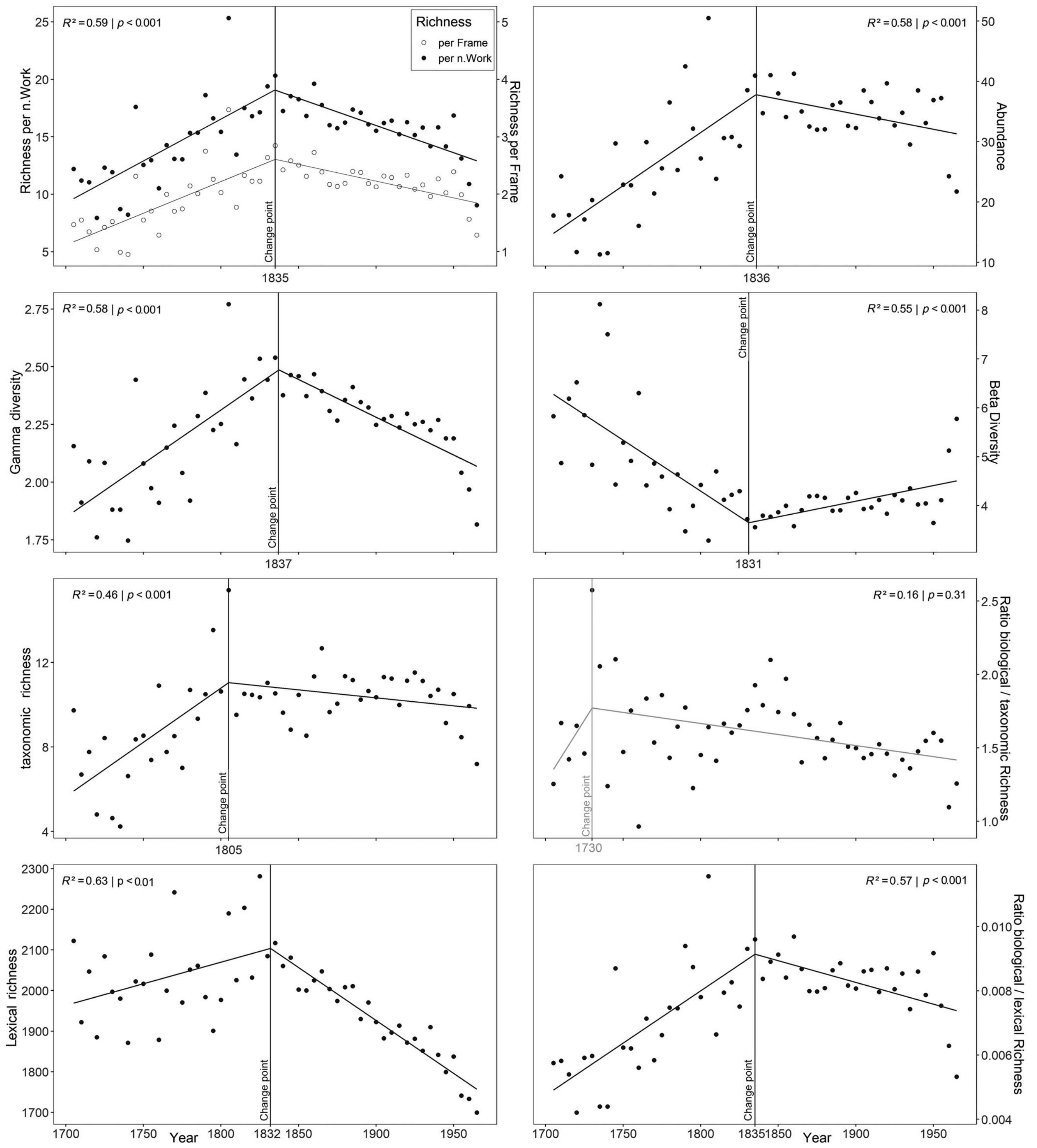
middle of the 20th century. To shed light on the potential causes for this decline we explored two potential drivers that are unrelated to biodiversity awareness: a general decline in lexical richness and systematic language streamlining. We found that the richness of taxon labels declined faster than general lexical richness. Furthermore, the lack of a significant trend in the richness of synonyms indicates there was little effect of streamlining on the biodiversity vocabulary used.

If we accept that changes in communication patterns may serve as indicators for cultural (Michel et al., 2011) and consequently societal change, these findings make our hypothetical scenario more probable. This scenario stated that biodiversity may progressively disappear from creative works towards modern times as a consequence of the declining exposure of people (and authors) to nature and its biodiversity in their daily lives because of industrialisation, urbanisation and intensified land use (Seppelt & Cumming, 2016). We point out though that as we lack information on important driver variables quantifying, for example, the change in the percentage of authors being raised and living in cities or historical changes in the social function of creative literature, we cannot claim a causal relationship, but nevertheless we can observe a historic co-occurrence between decreasing BiL and increasing industrialisation/urbanisation. In the following section we will detail how the initial increase and subsequent decline of BiL might correspond to our hypothetical scenario. Finally, we will discuss the limits and potential of our approach, including its relation to NCP.

4.1 | The rise of BiL in the 18th century

There are several possible reasons why biodiversity indicators might be expected to exhibit a general increase in BiL throughout the first half of the investigated period. The temporal coverage of the corpus includes the thematic and formal emergence of fiction during the 18th century, especially the rise of the novel as the paradigm of modern literature (Davis, 1983; Watt, 1957). The beginning of our study period therefore reflects the time before the onset of fiction literature, when BiL was rather low. We want to highlight that the Google Books corpus (M. Davies, 2011), albeit overwhelming in extent, contains zero to five fiction works annually between 1,700 and 1,720 and even fewer before. This confirms that (a) there was comparatively little fiction literature published in the first decades of the investigated period and that (b) above all, current access to the remainder in digitised form is low.

As well as being a generally small corpus in that time, the function of literature at the beginning of our study period was primarily to reflect and mostly confirm religious, moral, political and cultural



norms. With its emancipation from this normative function, more books were published and literature successively developed into a broader representation of people's communication and thereby increasingly also covered topics related to the surrounding biological environment. Combining this with our argumentation above, the diachronic increase in BiL appears to reflect at least two related transformations in society: On the one hand, an increase in biodiversity knowledge and awareness; on the other hand, a widening of the spectrum of possibilities within fictional literature. Considering these contexts, it is clear that the development of biodiversity in the first decades of our study is not solely a reflection of biodiversity within society's communication and mentality but is the result of several drivers, discussed below.

The 18th century remained a time for exploring unknown parts of the world and encountering flora and fauna unknown to the European world. Additionally, global consolidation of European

FIGURE 7 The development of biodiversity and lexical parameters for normalised works between 1705 and 1969. A normalised work is a randomised, one hundred times iterated subsample of ten 1,000-word-frames per work. For each 5-year period, we averaged the works by subsampling 10 of the works from that period by different authors and iterated this process one hundred times. We fitted a change point regression to the data, which provides an overall R^2 and p-value, an estimation of the change point and a linear model on either side as depicted by the regression lines. Insignificant regression lines are greyed out. Top left, the number of unique taxon labels (richness) per normalised work, and for comparison per 1,000-word-frame. For example between 1830 and 1834 a work, normalised to 10,000 words/10 frames, contained on average 19 different common names for biological taxa. Top right, the abundance of taxon labels. Upper centre left, the γ-diversity, as indicated by the Shannon–Weaver Index for occurrences of taxon labels within a normalised work. Upper centre right, the dissimilarity between frames as indicated by the β-diversity based on the Shannon–Weaver Index and Whittaker (1960) concerning the occurrences of taxon labels within a frame on average and the corresponding normalised work. Lower centre left, the taxonomic richness. We show the biologically distinct taxa, omitting any kind of synonymy, normalised within a 5-year period. Lower centre right, the richness of taxon labels (top left) relative to the taxonomic richness (lower centre right). We show the ratio between biological and taxonomic richness. Bottom left, the general lexical richness. We show the vocabulary; that is, the number of unique tokens (types) per normalised work. We utilised the lexical richness to calculate the relative use of taxon labels in relation to the overall trend of the size of the vocabulary. Bottom right, the richness of taxon labels (top left) relative to the lexical richness (bottom left). We show the ratio between biological and lexical richness

civilisations by imperialism and colonialism involved a broader information and knowledge transfer from cultures previously less known in Europe. This is likely to have induced a general awareness of the diversity of the world and promoted a desire for rich and adventurous tales in authors and audience alike, as can be seen in the rising popularity of travelogues (Rennie, 1995) or, for example, in Daniel Defoe's *Robinson Crusoe* (1719), which initiated a whole genre of adventurous island-narratives filled with plants, animals and ecologies (Armstrong, 2008, pp. 4–48). This may have further led to an increase in taxon labels in the literature of the 18th century.

The 18th century was also the Age of Enlightenment, which supported the distribution of worldly knowledge and its tendential preference over religious and moral content. This process promoted the expansion of a comprehensible educational and scholastic system (Fischer & Withers, 2021; Schaffar, 2014). Simultaneously, science disciplines diversified into specialised branches, including biology, allowing for a better understanding of the biological world (Bühler, 2016, Chapter 1; Toepfer, 2011). Together, these developments produced a general extension of the knowledge base, also enhancing the distribution of knowledge about the biosphere. It seems likely that this knowledge might have been another driver for biological diversification within literature.

We observed that the peak in BiL occurs after the onset of industrialisation around 1,780 (Coopersmith & Trebilcock, 1984). We assume that in the late 18th and in the beginning of the 19th centuries the proposed disconnecting mechanisms of urbanisation, industrialisation and land-use change (Grigg, 1987; Seppelt & Cumming, 2016) already diverted people's awareness away from nature. Since cultural movements in visual arts, literature and philosophy, such as romanticism and transcendentalism, can be understood as a reaction to the physical distancing from nature (Bate, 1991; Clark, 2011; J. Davies, 2018; Schläger, 1989), reflecting proto-ecological aspirations, they may have prolonged and initially counteracted the decrease in BiL. Additionally, people who grew up in a rural environment can be assumed to partially retain their knowledge, awareness and attitude towards nature after moving into urban areas and even transfer some of it to the next generation. By implication, this leads to a slower alienation process, which may have caused the delayed maximum of BiL.

Finally, we should expect an overall increase in BiL due to an increase in the detection rate of taxon labels towards the present time, as our list is generated based on contemporary word lists (Wikispecies and Wikidata) and may miss historical vernacular names, especially their regional variants.

4.2 | The fall of BiL after the 1830s—Novel insights from a comprehensive approach

The development of richness, on the level of both frame and normalised work, and γ-diversity shows distinct peaks in the 1830s, followed by downward trends. We speculate that industrialisation, urbanisation and land-use change initiated a disconnect between people and their natural environment, depriving them successively of their in-depth knowledge about, and awareness for, flora and fauna in their vicinity. Our rationale is that the daily lives of a large proportion of the population changed dramatically in the first half of the 19th century (Brown & Harrison, 1978), leading to a decrease in their exposure to nature, facilitated by factors like the physical distance between natural and urban areas, by the growing employment in factories and manufactures as opposed to nature-bound professions, the decreasing necessity of knowing how to work with nature or even to survive in the wild and by denaturalisation of formerly pristine landscapes (Trepl, 1987). In section 4.1, we argued for several synchronously acting drivers for the increase in the first half of the investigated period and we expect that the joint influence of these positive drivers extends well beyond the peak of BiL in the middle of the 19th century. A decline of BiL after 1835 may thus require a counterforce, one that not only compensates for the trend before the 1830s but is even strong enough to reverse it. This overlap of upward and downward trends may also have contributed to the delay of the peak discussed in Section 4.1.

The development of abundance, showing only a slow decrease after the 1830s, indicates that authors still valued flora and fauna and integrated it into their plot settings. It follows that the decrease in BiL was not a result of decreasing interest or a decline in nature-based subjects, as exhibited and confirmed in movements like naturalism and realism. Counterintuitively, the poetological attitudes of these periods, according to our results, do not induce a strong return to BiL and instead appear to be characterised by a tendency towards lower biodiversity than during the period of romanticism. Also, countermovements to romanticism turning away from a nature-appraising attitude have been reported, sometimes by bemoaning the loss of nature (Jauss, 1989), but mostly by focussing on political issues, social injustice and on the challenges of modernisation, such as urbanisation and industrialisation. Writing in a 'realistic' or 'naturalistic' manner was, increasingly, writing about human concerns in cultured spaces.

The β-diversity is a rough indicator of the biological dissimilarity of sections within a single work, the calculation of which requires a comprehensive list of taxon labels like the one assembled for this investigation. We observe it to be nearly constant after reaching its minimum simultaneously with the maximum of γ-diversity. From this development, we conclude that in the beginning of the 18th century the design of the ensembles of animals and plants between individual scenes and settings was more diverse than in later phases of the investigated period. The constant and relatively uniform presence of taxon labels from about 1,830 onwards strengthens our earlier argument that the decline of BiL was not a result of declining interest in nature. Instead, it appears that authors both grew unable to distinguish the biodiversity between specific real-life environments and exhibited a tendency to homogenise animals and plants for different settings within a work. Because of these insights, we assume β-diversity to be another meaningful indicator for the decline of biodiversity awareness we presumed above.

Our analysis of the general lexical richness indicates a streamlining process of the English language after the 1830s. Therefore, we propose that the development in the second half of the investigated period is not only driven by decreasing detailed knowledge about and awareness of the multitude of living things in the immediate environment, but also by decreasing expressions of diversity as a consequence of streamlining. We surmise that this is a process mainly driven by standardisation of language, including (a) the conflation of synonyms, (b) the standardisation of orthography, (c) a de-regionalisation trend and (d) general obsolescence of supposedly dispensable vocabulary. While this hypothesis needs to be further investigated, the development of lexical richness clearly shows that since the 1830s a larger overall vocabulary was lost rather than gained by potentially enriching processes like globalisation (e.g. colonisation) or technological (e.g. electricity, telephone), economic (e.g. material prosperity) and social (e.g. improving education) progression (Brown & Harrison, 1978). Although clearly beyond the scope of this investigation, this indicates that our language is not only subject to the loss of living beings in our immediate environment, but also to a depletion within other realms of human life, possibly including the labour market and social structures. The development of the biological, relative to the lexical richness, shows a slight but still significant decrease after the 1830s. Also beyond the scope of this investigation, the development in the past 50 years should be investigated separately with a suitable corpus, keeping the diverse drivers during the onset of the internet in the 1990s in mind.

We also analysed the development of distinguished biological taxa mentioned in the corpus, as opposed to labels for taxa, allowing us to determine the development of synonymy for biological entities. We found that there is no significant streamlining effect within the vernacular biological vocabulary. The decreasing ratio of taxon label to lexical richness as well as the exclusion of a discernible streamlining within the biological vocabulary illustrates the prominence of the decrease in BiL over general streamlining in addition to its overall decline and supports our argumentation towards declining biodiversity awareness.

We would like to point out that our analysis of the general lexical richness only refers to the English language and there may be differing trends in other languages due to language policies, cultural attitudes and shifts in cultural dominance in the global or regional context. Translations of non-English works also do not appropriately reflect the variety of the language of origin, but also the characteristics of the English language at the time of translation, the proficiency of the translator in both languages and the translatability between the languages. As an example, our corpus contains a vocabulary of barely 30,000 words for the German author Johann Wolfgang von Goethe, despite the fact that his German vocabulary was around 100,000 words (Eisenberg, 2013, Chapter 1). Our corpus contains only about one quarter of his œuvre, however, this alone cannot explain such a large discrepancy, as the inclusion of additional works adds successively fewer new words to an author's total vocabulary. The vocabulary for animals and plants is of a rather technical nature, as each term is intended to refer to a distinct taxon. Additionally, translations tend to preserve the basic nature of entities within books, which is why we do not expect a significant difference between animals or plants mentioned. This may only lead to a negligible difference of BiL between the original language and its English translation.

Authors tend to embed their creation into concepts the reader is assumed to be familiar with, meaning that the placement of animals and plants within literature not only reflects the authors' own but also the assumed average audiences' awareness for biological diversity (Eckert & Stacey, 2000; Wolff et al., 1999). Moreover, creative literature captures metaphorical and symbolic meaning, which usually originated because of properties associated with the respective animal or plant in the first place. We argue that these two properties, (a) awareness for and assessment of the natural environment of author and readership as representatives of the society, and (b) coverage of metaphorical meanings, contribute to exposing the characteristics of a society's mentality and thus correlate with literary biodiversity awareness.

Various studies report there is a close relationship between our environment and our cultural attitudes (Richards & Tunçer, 2018; Willemen et al., 2015) as well as between our cultural attitudes and our communication (Michel et al., 2011) and minds (Wolff et al., 1999). Therefore, with our results showing an actual decline in BiL, we argue that the loss of biodiversity or exposure to it in our daily lives may impoverish our thinking patterns, for instance, our understanding of mechanisms and structures in living nature. This may deprive us of a sophisticated code constructed from natural examples and eventually of cultural references. As a consequence, we may possibly experience detrimental effects on our creativity, ability to communicate and mental health—all components of human wellbeing (Celume et al., 2017; Conner et al., 2018; Segrin, 2005). Our study presents the development of BiL within our communication, which can be seen as the documentation of how humans pass on thoughts. In this line, we concur with Wolff et al. (1999) that the diversity of cultural products, especially the diversity of language in written communication, reflects the diversity of mental processes. Therefore, we hypothesise that the ascertained use of taxon labels within our communication is correlated with the synchronic biodiversity awareness of Western society with the respective time period, as we further argue below.

BiL provides a glimpse into some preliminary parameters concerning NCP. We cannot directly translate BiL into biodiversity awareness or into a quantification of NCP with this investigation alone. However, we have revealed previously overlooked connections between living nature and cultural products. With our determination of the mean percentage of taxon labels within literature, we have shown that there seems to be a certain necessity or demand, respectively, for biodiversity, of approximately 1% in Western creative literature according to our results, within cultural products. We therefore encourage dialogue between scholars of the humanities and the natural sciences to facilitate a better understanding of the entanglement of nature and culture. In this light, our results need to be assessed in terms of their actual contribution to human wellbeing. Such an assessment could be promoted by an investigation of the development of society's attitudes towards the individual animals and plants identified within our investigation, applying, for example, sentiment analysis (Shanahan et al., 2006), topic modelling (Blei et al., 2003) and further integration of literary scholarship from the environmental humanities. A sentiment analysis could provide insights into the positioning of works in the spectrum between partially unrealistic, positive and negative attitudes towards the biological world.

4.3 | Further limitations, observations and the future potential of our approach

To our knowledge this is the first study using such comprehensive data resources as well as the first investigation combining computational literary studies with biodiversity indexes of ecological theory. However, there is still a lack of relevant data and related research, which is why we made several assumptions. We estimated the year of publication by using a shifted mean of the years of birth and death of the corresponding author. This resulted in the placement of all works by one author into a single 5-year interval. As authors have their own idiosyncratic style of writing and awareness for biodiversity, this placement influenced the data point of one interval strongly, whereas their influence does not extend to the adjacent intervals. Especially when fewer works are available, this resulted in a high variation of the individual data points as well as in a slightly disproportionate reflection of the authors' importance, influence and individual development throughout their lifetime. However, this has only a minor effect on the overall trend, as the standard deviation of 6.9 years against the manually tested subsample is smaller than the width of two intervals. The high amount of data as well as the subsampling method for each interval produces reliable and minimally biased average values.

This study covers all literary forms (e.g. novels, plays, short story collections, poetry collections) and genres (e.g. adventure, social, fantasy, travelogue), but unambiguous categorisation in this regard is not provided within the metadata of the corpus. As these categories may strongly influence the overall development, further investigations with more detailed metadata are needed. Children's books, for example, tend to contain a comparatively large proportion of world knowledge for educational purposes and thus may exhibit a different individual trend to that of other genres (Varga, 2009). For future investigations we suggest increasing the pool of metadata by combining manual and automated methods of information acquisition.

Due to the norms of literary discourse, which demands originality rather than mainstreaming, authors developed their individual styles of writing, as mentioned above. As a consequence, they differ greatly in their taxon label vocabulary. For example, novelist Charlotte Mary Yonge's (1823–1901) high taxon label content may be the result of her general tendency towards educational intention in her books and of her favourite subjects being history and biology. Both Jules Verne (1828–1905) and Robert William Chambers (1865–1933) produced a variety of fantastic fiction works, thereby contriving a multitude of different settings with varying biodiversity ensembles. Both authors used more than 700 different taxon labels in their works contained within Project Gutenberg. Apart from fiction works, the corpus also contains other creative literature like travelogues or biographies. Robert Kerr (1757–1813) and James George Frazer (1854–1941) are authors with a high taxon label richness within their works outside fiction literature. A prominent reason for this is that both authors drew inspiration from a range of sources: the former described a number of voyages of navigation, commerce and discovery, and the latter published several works about the progression of mythology and beliefs. Among the authors in the corpus, there is no tendency towards a saturating taxon label richness with larger vocabulary, showing that authors attempt to enrich their scenes with biodiversity according to the extent of their œuvre and their level of lexical richness. We point out that although most authors with high taxon label richness published their works after the 1830s, BiL declined according to our analysis. This indicates that around the 1830s taxon labels were more uniformly and extensively present.

As a general tendency, we observed that the proportion of domesticated animals and plants among the one hundred most abundant taxon labels is higher than their proportion within the taxonomic classification. This preference for useful and edible species is not surprising as they play a significant role in humans' daily lives. With a shift from rural to urban settings, the decrease in BiL we observed could thus partly reflect a decreasing reference to pets, livestock and crops. In order to explore this, we would need additional metadata regarding the domestic use of species behind our taxon labels, which we plan for a follow-up study. Similarly, we could advance our understanding of the decrease in BiL by (a) extending our metadata to indicate the exotic origins of species and (b) applying taxonomic data to the occurrences to indicate deviations in the influence between different taxa, like birds, butterflies or flowers.

Because of the large number of individual contributors, openly available sources potentially contain errors and deviations in the process of providing and formatting the data, which is why we applied methods for cleaning and error reduction. Although we cannot guarantee that we found every individual taxon label within the corpus without any false positives, we argue that our comprehensive approach is likely to reveal the relevant patterns of use of taxon labels. In contrast, approaches using limited taxon label lists (Kesebir & Kesebir, 2017; Queiroz et al., 2015) can be misleading, as they may ignore potentially compensating trends exhibited by labels not covered by the corresponding list. In several cases (Ladle et al., 2016; Proulx et al., 2014) such investigations have been carried out using the Google Books (Davies, 2011) corpus accessed via the Google Ngram Viewer (Lin et al., 2012). In addition, our comprehensive approach allowed us to go beyond simple word counts by quantifying different indicative facets of BiL, including the Shannon diversity and the corresponding β-diversity within a work normalised to 10,000 words.

In the course of this study several points have been made that suggest a strong correlation of BiL to the biodiversity awareness of the society at the respective time. This novel approach required a corpus with individual works, which in future research projects would also allow us to include specific metadata on books and authors in our analysis. The Google Ngram Viewer does not allow for such a comprehensible and size-normalised investigation to be carried out and is not suitable for correlating trends to specific characteristics of authors or works. Other open fiction corpora, such as the Corpus of Historical American English (M. Davies, 2010), British National Corpus (Burnage & Dunlop, 1992) or Canon of Western Literature (Green, 2017), were either smaller, typically did not cover the preindustrial era for comparison or partially contained a specific preselected canon, not allowing for random sampling to represent a balanced cross section through the diverse literature.

5 | CONCLUSION AND PROSPECT

Our project developed a novel method to quantify BiL, thereby potentially pointing towards society's biodiversity awareness at the respective time as reflected in fiction literature. Our hypothesis of declining BiL within the used corpus due to or temporally associated with industrialisation, land-use change and urbanisation was confirmed by the data and we could locate the start of this effect in the 1830s. From our analysis, we conclude that this negative trend of BiL may reflect the detachment from nature progressively dominating over the expansion of education and natural sciences and exceeding the effect of overall language streamlining.

We regard this finding as relevant for the assessment and interpretation of nature's non-material contributions to people (Díaz et al., 2015; Pascual et al., 2017). Our approach to quantify BiL has the potential to be used and interpreted in multiple ways and could be applied to a wide range of text genres (e.g. newspaper articles, twitter feeds, etc.). This would allow us to trace changes in general inclinations, sentiments, knowledge and awareness (Willemen et al., 2015; Wolff et al., 1999). Furthermore, this could then help us to analyse the role of biodiversity for recreation, education and ultimately human wellbeing (Ladle et al., 2016) by the effect of nature's contribution to communication. We suggest that decision makers consider our approach as one tool to evaluate the effectiveness of biodiversity policies.

With this project, we raised a number of continuative questions. To make a strong case, future research, facilitated by the assembly of further metadata, should concentrate on answering more specific questions about works, authors and living things to unravel causal relationships and any mechanistic underpinning as well as to elucidate regional or genre-specific differentiation of the patterns revealed in this work.

ACKNOWLEDGEMENTS

We express our thanks to W. P. Klein, Esther Köhring, Liza Bauer, Frederike Middelhoff, Marcela Hernández, Paul Sladen and Bryan Norton for their valuable feedback and insights in the field of literary studies. We thank N. Kowalewski and F. Küchler, who helped to collect large amounts of metadata. We also thank M. Beckmann for his essential inspiration and feedback. We thank Alfons van der Plas, Kathryn Barry, Claas-Thido Pfaff and Ronny Richter for their constant support and feedback throughout the project. This research was also supported by the Helmholtz Centre for Environmental Research. We acknowledge support from Leipzig University for Open Access Publishing.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHORS' CONTRIBUTIONS

C.W. conceived the project idea and acquired funding; L.L., C.W. and R.S. conceptualised the study; L.L., C.W. and M.B. developed the methodology; L.L. gathered and processed the data and conducted the search and the analysis; L.L., C.W., M.B., R.B. and K.B. interpreted the results; L.L. wrote the first draft of the manuscript. All authors contributed substantially to the writing of the manuscript.

DATA AVAILABILITY STATEMENT

The Standardized Project Gutenberg Corpus used in this study is available in the repository pgcorpus/Gutenberg at GitHub (https://

github.com/pgcorpus/gutenberg) or at Zenodo (https://zenodo.org/record/2422561) as a permanent record (Gerlach & Font-Clos, 2018) and was reduced using a list of key words (see Appendices A1 and A2) as described in section 2.1. The data of Wikidata and Wikispecies used to generate the database of taxon labels and part of the metadata for authors and their works in this study are available as database dumps from Wikimedia (https://dumps.wikimedia.org).

ORCID

Lars Langer https://orcid.org/0000-0002-1076-2936
Manuel Burghardt https://orcid.org/0000-0003-1354-9089
Katrin Böhning-Gaese https://orcid.org/0000-0003-0477-5586
Ralf Seppelt https://orcid.org/0000-0002-2723-7150
Christian Wirth https://orcid.org/0000-0003-2604-8056

REFERENCES

Ainscough, J., de Vries Lentsch, A., Metzger, M., Rounsevell, M., Schröter, M., Delbaere, B., de Groot, R., & Staes, J. (2019). Navigating pluralism: Understanding perceptions of the ecosystem services concept. *Ecosystem Services*, *36*. https://doi.org/10.1016/j.ecoser.2019.01.004

Armstrong, P. (2008). *What animals mean in the fiction of modernity*. Routledge.

Atran, S., Estin, P., Coley, J. D., & Medin, D. L. (1997). Generic species and basic levels: Essence and appearance in folk biology. *Journal of Ethnobiology*, *17*, 17–43.

Atran, S., & Medin, D. L. (2008). *The native mind and the cultural construction of nature*. https://doi.org/10.7551/mitpress/7683.001.0001

Barad, K. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Duke University Press.

Bate, J. (1991). *Romantic ecology: Wordsworth and the environmental tradition*. Routledge Revivals.

Biraben, J.-N. (2003). L'évolution du nombre des hommes. *Population et Sociétés*, *394*, 1–4.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Brown, D., & Harrison, M. J. (1978). *A sociology of industrialisation: An introduction*. https://doi.org/10.1007/978-1-349-15924-6

Bühler, B. (2016). *Ecocriticism*. https://doi.org/10.1007/978-3-476-05489-0

Burnage, G., & Dunlop, D. (1992). Encoding the British National Corpus. In J. Aarts, P. de Haan, & N. Oostdijk (Eds.), *English language corpora: Design, analysis and exploitation* (pp. 79–95). Rodopi.

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., Narwani, A., Mace, G. M., Tilman, D., Wardle, D. A., Kinzig, A. P., Daily, G. C., Loreau, M., Grace, J. B., Larigauderie, A., Srivastava, D. S., & Naeem, S. (2012). Biodiversity loss and its impact on humanity. *Nature*, *486*(7401), 59–67. https://doi.org/10.1038/nature11148

Celis-Diez, J. L., Díaz-Forestier, J., Márquez-García, M., Lazzarino, S., Rozzi, R., & Armesto, J. J. (2016). Biodiversity knowledge loss in children's books and textbooks. *Frontiers in Ecology and the Environment*, *14*(8), 408–410. https://doi.org/10.1002/fee.1324

Celume, M.-P., Sovet, L., Lubart, T., & Zenasni, F. (2017). The relationship between children's creativity and well-being at school. In M.-P. Celume, L. Sovet, T. Lubart, F. Zenasni, & F. K. Reisman (Eds.), *Creativity, innovation and wellbeing* (pp. 346–362). KIE Conference Publications.

Clark, T. (2011). Old world romanticism. In *The Cambridge introduction to literature and the environment* (pp. 15–24). https://doi.org/10.1017/CBO9780511976261.004

Cole, E. W. (1879). *Cole's funny picture book no. 1*. Cole Publications.

Conner, T. S., DeYoung, C. G., & Silvia, P. J. (2018). Everyday creative activity as a path to flourishing. *The Journal of Positive Psychology*, *13*(2), 181–189. https://doi.org/10.1080/17439760.2016.1257049

Coopersmith, J. C., & Trebilcock, C. (1984). The industrialization of the continental powers 1780–1914. *Technology and Culture*, *25*(2), 341. https://doi.org/10.2307/3104737

Daily, G. C. (2000). Ecology: The value of nature and the nature of value. *Science*, *289*(5478), 395–396. https://doi.org/10.1126/science.289.5478.395

Daniel, T. C., Muhar, A., Arnberger, A., Aznar, O., Boyd, J. W., Chan, K. M. A., Costanza, R., Elmqvist, T., Flint, C. G., Gobster, P. H., Gret-Regamey, A., Lave, R., Muhar, S., Penker, M., Ribe, R. G., Schauppenlehner, T., Sikor, T., Soloviy, I., Spierenburg, M., … von der Dunk, A. (2012). Contributions of cultural services to the ecosystem services agenda. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(23), 8812–8819. https://doi.org/10.1073/pnas.1114773109

Davies, J. (2018). Romantic ecocriticism: History and prospects. *Literature Compass*, *15*(9). https://doi.org/10.1111/lic3.12489

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, *25*(4), 447–464. https://doi.org/10.1093/llc/fqq018

Davies, M. (2011). *Google books (American English) Corpus (155 billion words, 1810–2009)*. Retrieved from http://googlebooks.byu.edu/

Davis, L. J. (1983). *Factual fictions: The origins of the English novel*. Columbia University Press.

de Araujo Barbosa, C. C., Atkinson, P. M., & Dearing, J. A. (2015). Remote sensing of ecosystem services: A systematic review. *Ecological Indicators*, *52*, 430–443. https://doi.org/10.1016/j.ecolind.2015.01.007

Defoe, D. (1719). *The life and strange surprizing adventures of Robinson Crusoe*. William Taylor.

Detering, H. (2020). *Menschen im Weltgarten - Die Entdeckung der Ökologie in der Literatur von Haller bis Humboldt*. Wallstein.

Díaz, S., Demissew, S., Carabias, J., Joly, C., Lonsdale, M., Ash, N., Larigauderie, A., Adhikari, J. R., Arico, S., Báldi, A., Bartuska, A., Baste, I. A., Bilgin, A., Brondizio, E., Chan, K. M. A., Figueroa, V. E., Duraiappah, A., Fischer, M., Hill, R., … Zlatanova, D. (2015). The IPBES Conceptual Framework – Connecting nature and people. *Current Opinion in Environmental Sustainability*, *14*, 1–16. https://doi.org/10.1016/j.cosust.2014.11.002

Díaz, S., Pascual, U., Stenseke, M., Martín-López, B., Watson, R. T., Molnár, Z., … Shirayama, Y. (2018). Assessing nature's contributions to people. *Science*, *359*(6373), 270–272. https://doi.org/10.1126/science.aap8826

Eckert, C., & Stacey, M. (2000). Sources of inspiration: A language of design. *Design Studies*, *21*(5), 523–538. https://doi.org/10.1016/S0142-694X(00)00022-3

Eisenberg, P. (2013). *Grundriss der deutschen Grammatik, Band 1 - Das Wort*. N. Fuhrhop (Ed.). https://doi.org/10.1007/978-3-476-00757-5

Fischer, L., & Withers, C. W. J. (2021). Geographical education in the eighteenth-century German-speaking territories. *Paedagogica Historica*, 1–27. https://doi.org/10.1080/00309230.2021.1872658

Gagliano, M., Ryan, J. C., & Vieira, P. (2017). *The language of plants: Science, philosophy, literature*. University of Minnesota Press.

Gerlach, M., & Font-Clos, F. (2018). Standardized project Gutenberg Corpus. https://doi.org/10.5281/zenodo.2422561

Gerlach, M., & Font-Clos, F. (2020). A standardized project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, *22*(1), 126. https://doi.org/10.3390/e22010126

Goodrich, S. G. (1845). *Illustrative anecdotes of the animal kingdom*. Bradbury, Soden & Co.

Green, C. (2017). Introducing the corpus of the canon of western literature: A corpus for culturomics and stylistics. *Language and Literature: International Journal of Stylistics*, *26*(4), 282–299. https://doi.org/10.1177/0963947017718996

Greenaway, C. (1884). *Language of flowers*. Routledge.

Grigg, D. B. (1987). The industrial revolution and land transformation. In M. G. Wolman & F. G. A. Fournier (Eds.), *Land transformation in agriculture* (pp. 79–109). Wiley-Blackwell.

Haraway, D. J. (2016). Staying with the trouble. In *Staying with the trouble*. https://doi.org/10.1215/9780822373780

Haraway, D. J., & Begelke, M. (2003). *The companion species manifesto: Dogs, people, and significant otherness*. Prickly Paradigm Press.

Hudson, N. (2015). Literature and social class in the eighteenth century. *Oxford handbooks online*, https://doi.org/10.1093/oxfordhb/9780199935338.013.007

IPBES. (2019). *Summary for policymakers of the global assessment report on biodiversity and ecosystem services* (In S. Díaz, J. Settele, E. S. Brondízio, H. T. Ngo, M. Guèze, J. Agard, A. Arneth, P. Balvanera, K. A. Brauman, S. H. M. Butchart, K. M. A. Chan, L. A. Garibaldi, K. Ichii, J. Liu, S. M. Subramanian, G. F. Midgley, P. Miloslavich, Z. Molnár, D. Obura, A. Pfaff, S. Polasky, A. Purvis, J. Razzaque, B. Reyers, R. R. Chowdhury, Y. J. Shin, I. J. Visseren-Hamakers, K. J. Willis, & C. N. Zayas (Eds.)]. IPBES Secretariat. https://doi.org/10.5281/zenodo.3553579

Jauss, H.-R. (1989). Ursprünge der Naturfeindschaft in der Ästhetik der Moderne. In H.-D. Weber (Ed.), *Vom Wandel des neuzeitlichen Naturbegriffs. Konstanzer Bibliothek Bd. 13* (pp. 207–225). Konstanzer Universitätsverlag.

Kesebir, S., & Kesebir, P. (2017). A growing disconnection from nature is evident in cultural products. *Perspectives on Psychological Science*, *12*(2), 258–269. https://doi.org/10.1177/1745691616662473

Krausmann, F., & Haberl, H. (2002). The process of industrialization from the perspective of energetic metabolism. *Ecological Economics*, *41*(2), 177–201. https://doi.org/10.1016/S0921-8009(02)00032-0

Ladle, R. J., Correia, R. A., Do, Y., Joo, G.-J., Malhado, A. C. M., Proulx, R., Roberge, J.-M., & Jepson, P. (2016). Conservation culturomics. *Frontiers in Ecology and the Environment*, *14*(5), 269–275. https://doi.org/10.1002/fee.1260

Lambin, E. F., & Geist, H. (2006). *Land-use and land-cover change*. E. F. Lambin & H. Geist (Eds.). https://doi.org/10.1007/3-540-32202-7

Lautenbach, S., Mupepele, A.-C., Dormann, C. F., Lee, H., Schmidt, S., Scholte, S. S. K., Seppelt, R., van Teeffelen, A. J. A., Verhagen, W., & Volk, M. (2019). Blind spots in ecosystem services research and challenges for implementation. *Regional Environmental Change*, *19*(8), 2151–2172. https://doi.org/10.1007/s10113-018-1457-9

Lautenbach, S., Seppelt, R., Liebscher, J., & Dormann, C. F. (2012). Spatial and temporal trends of global pollination benefit. *PLoS ONE*, *7*(4). https://doi.org/10.1371/journal.pone.0035954

Lin, Y., Michel, J.-B., Aiden, E., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the Google Books Ngram Corpus. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Long papers* (Vol. 1). Retrieved from https://www.aclweb.org/anthology/P12-1

Maes, J., Fabrega, N., Zulian, G., Barbosa, A. L., Vizcaino, P., Ivits, E., … Lavalle, C. (2015). *Mapping and assessment of ecosystems and their services: Trends in ecosystems and ecosystem services in the European Union between 2000 and 2010*. https://doi.org/10.2788/341839

Magurran, A. E., & McGill, B. J. (2011). *Biological diversity: Frontiers in measurement and assessment*. Oxford University Press.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: System demonstrations* (pp. 55–60). https://doi.org/10.3115/v1/P14-5010

Martinez-Alier, J. (2002). *The environmentalism of the poor*. https://doi.org/10.4337/9781843765486

McCrindle, C. M. E., & Odendaal, J. S. J. (1994). Animals in books used for preschool children. *Anthrozoös*, *7*(2), 135–146. https://doi.org/10.2752/089279394787001998

Mesoudi, A. (2011). *Cultural evolution: How Darwinian theory can explain human culture and synthesize the social sciences*. University of Chicago Press.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., … Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182. https://doi.org/10.1126/science.1199644

Millennium Ecosystem Assessment. (2005). In W. V. Reid, H. A. Mooney, A. Cropper, D. Capistrano, S. R. Carpenter, K. Chopra, … M. B. Zurek (Eds.), *Ecosystems and human well-being: Synthesis*. Island Press.

Morton, T. (2007). *Ecology without nature: Rethinking environmental aesthetics*. Harvard University Press.

Pascual, U., Balvanera, P., Díaz, S., Pataki, G., Roth, E., Stenseke, M., Watson, R. T., Başak Dessane, E., Islar, M., Kelemen, E., Maris, V., Quaas, M., Subramanian, S. M., Wittmer, H., Adlan, A., Ahn, S. E., Al-Hafedh, Y. S., Amankwah, E., Asah, S. T., … Yagi, N. (2017). Valuing Nature's Contributions to People: The IPBES approach. *Current Opinion in Environmental Sustainability*, *26–27*, 7–16. https://doi.org/10.1016/j.cosust.2016.12.006

Prévot-Julliard, A.-C., Julliard, R., & Clayton, S. (2015). Historical evidence for nature disconnection in a 70-year time series of Disney animated films. *Public Understanding of Science*, *24*(6), 672–680. https://doi.org/10.1177/0963662513519042

Proulx, R., Massicotte, P., & Pépino, M. (2014). Googling trends in conservation biology. *Conservation Biology*, *28*(1), 44–51. https://doi.org/10.1111/cobi.12131

Queiroz, A. I., Fernandes, M. L., & Soares, F. (2015). The Portuguese literary wolf. *Digital Scholarship in the Humanities*, *30*(3), 388–404. https://doi.org/10.1093/llc/fqt069

Rennie, N. (1995). *Far-fetched facts: The literature of travel and the idea of the South Seas*. Clarendon Press.

Richards, D. R., & Tunçer, B. (2018). Using image recognition to automate assessment of cultural ecosystem services from social media photographs. *Ecosystem Services*, *31*, 318–325. https://doi.org/10.1016/j.ecoser.2017.09.004

Rigby, K. (2014). Romanticism and ecocriticism. In G. Garrard (Ed.), *The Oxford handbook of ecocriticism* (Vol. 1, pp. 60–79). https://doi.org/10.1093/oxfordhb/9780199742929.013.003

Rigby, K. (2020). *Reclaiming romanticism – Towards an ecopoetics of decolonization*. Bloomsbury Academic.

Schaffar, B. (2014). Changing the definition of education. On Kant's educational paradox between freedom and restraint. *Studies in Philosophy and Education*, *33*(1), 5–21. https://doi.org/10.1007/s11217-013-9357-4

Schläger, J. (1989). Landschaft, Natur und Individualität in der englischen Romantik. In H.-D. Weber (Ed.), *Vom Wandel des neuzeitlichen Naturbegriffs. Konstanzer Bibliothek Bd. 13* (pp. 177–206). Konstanzer Universitätsverlag.

Schmid, B., Balvanera, P., Cardinale, B., Godbold, J., Pfisterer, A., Raffaelli, D., … Srivastava, D. (2009). Consequences of species loss for ecosystem functioning: Meta-analyses of data from biodiversity experiments. *Biodiversity, Ecosystem Functioning, and Human Wellbeing*, 14–29. https://doi.org/10.5167/uzh-25528

Schmidt, S., Manceur, A. M., & Seppelt, R. (2016). Uncertainty of monetary valued ecosystem services – Value transfer functions for global mapping. *PLoS ONE*, *11*(3). https://doi.org/10.1371/journal.pone.0148524

Segrin, C. (2005). Communication and the study of personal well-being. *Gazette (Leiden, Netherlands)*, *67*(6), 547–549. https://doi.org/10.1177/0016549205057549

Seppelt, R., & Cumming, G. S. (2016). Humanity's distance to nature: Time for environmental austerity? *Landscape Ecology*, *31*(8), 1645–1651. https://doi.org/10.1007/s10980-016-0423-5

Seppelt, R., Dormann, C. F., Eppink, F. V., Lautenbach, S., & Schmidt, S. (2011). A quantitative review of ecosystem service studies: Approaches, shortcomings and the road ahead. *Journal of Applied Ecology*, *48*(3), 630–636. https://doi.org/10.1111/j.1365-2664.2010.01952.x

Shanahan, J. G., Qu, Y., & Wiebe, J. (2006). *Computing attitude and affect in text: Theory and applications*. The Information Retrieval Series (Vol. 20). Springer.

Sumarga, E., Hein, L., Edens, B., & Suwarno, A. (2015). Mapping monetary values of ecosystem services in support of developing ecosystem accounts. *Ecosystem Services*, *12*. https://doi.org/10.1016/j.ecoser.2015.02.009

Tilman, D. (1999). The ecological consequences of changes in biodiversity: A search for general principles. *Ecology*, *80*(5), 1455. https://doi.org/10.2307/176540

Toepfer, G. (2011). *Historisches Wörterbuch der Biologie - Geschichte und Theorie der biologischen Grundbegriffe. Band 2: Gefühl–Organismus*. J.B. Metzler.

Trepl, L. (1987). *Geschichte der Ökologie. Vom 17. Jahrhundert bis zur Gegenwart*. https://doi.org/10.1002/mmnz.19880640216

Tüür, K., & Tønnessen, M. (2014). The semiotics of animal representations. In *The semiotics of animal representations* (pp. 7–30). https://doi.org/10.1163/9789401210720_002

Ulloa-Torrealba, Y., Stahlmann, R., Wegmann, M., & Koellner, T. (2020). Over 150 years of change: Object-oriented analysis of historical land cover in the main river catchment, Bavaria/Germany. *Remote Sensing*, *12*(24), 4048. https://doi.org/10.3390/rs12244048

Varga, D. (2009). Babes in the woods: Wilderness aesthetics in children's stories and toys, 1830–1915. *Society & Animals*, *17*(3), 187–205. https://doi.org/10.1163/156853009X445370

Vrandečić, D., & Krötzsch, M. (2014). Wikidata. *Communications of the ACM*, *57*(10), 78–85. https://doi.org/10.1145/2629489

Watt, I. P. (1957). *The rise of the novel: Studies in Defoe, Richardson and Fielding*. University of California Press.

Whittaker, R. H. (1960). Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs*, *30*(4), 407. https://doi.org/10.2307/1948435

Willemen, L., Cottam, A. J., Drakou, E. G., & Burgess, N. D. (2015). Using social media to measure the contribution of red list species to the nature-based tourism potential of African protected areas. *PLoS ONE*, *10*(6), e0129785. https://doi.org/10.1371/journal.pone.0129785

Wolff, P., Medin, D. L., & Pankratz, C. (1999). Evolution and devolution of folkbiological knowledge. *Cognition*, *73*(2), 177–204. https://doi.org/10.1016/S0010-0277(99)00051-7

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Langer, L., Burghardt, M., Borgards, R., Böhning-Gaese, K., Seppelt, R., & Wirth, C. (2021). 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*, 00, 1–17. https://doi.org/10.1002/pan3.10256