Deep distant reading
The rise of realism in Scandinavian literature as a case study
Bjerring-Hansen, Jens; Wilkens, Matthew

Published in:
Orbis Litterarum

DOI:
10.1111/oli.12396

Publication date:
2023

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Deep distant reading: The rise of realism in Scandinavian literature as a case study

Jens Bjerring-Hansen\textsuperscript{1} | Matthew Wilkens\textsuperscript{2}

\textsuperscript{1}University of Copenhagen, Copenhagen, Denmark
\textsuperscript{2}Cornell University, Ithaca, New York, USA

Correspondence
Jens Bjerring-Hansen, University of Copenhagen, Copenhagen, Denmark.
Email: jbh@hum.ku.dk

Funding information
The Carlsberg Foundation, Grant/Award Number: CF19-0661

Abstract
In this article we make a case for a synchronic and contextualizing perspective on the scaling of literary data, one which qualifies and expands the data points in terms of depth, or thickness, through the help of metadata on the social and historical conditions of the texts. Our case study is an investigation of the rise and impact of realism in a corpus of more than 800 Danish and Norwegian novels from 1870 to 1899 with three, interlocking critical and methodological aims. We use textual features to model realism in a large corpus of Danish-language novels. We compare the features driving that model to the ones that were important to the historical and critical development of realism as a literary project. And we use the results of our model to study the interplay between realism and social history. Our findings suggest that realism was more prominent and more widely distributed in Danish-language novels of the late nineteenth century than the critical tradition has usually acknowledged. Women appear to have written realist fiction not only at rates similar to men, but at times and with features that are difficult to distinguish from their male counterparts. The article relies on—and insists on further—dialogue between digital and analog approaches to the exploration of cultural data.

KEYWORDS
computational and quantitative literary studies, cultural
1 | INTRODUCTION

Reading at scale is the crux of a major strain of digital literary studies. Its main contributions to its mother discipline are the development, application, and discussion of computational techniques that allow for analysis of large text corpora in order to capture widespread and long-term literary trends. Consequently, most of the existing work on, and of, scale relies on a quantitative imperative governed by the ideals of more text (“big data”) and more time (longue durée). We pursue this imperative in conjunction with a more synchronic and contextualizing perspective, one that qualifies and expands textual data points in terms of depth, or thickness, through the help of metadata on the social and historical conditions under which those texts were produced.

This approach is not quite the same thing as a combination of close and distant reading, or what Martin Mueller (2020) has usefully called “scalable reading”. In that case, the aim is to move back and forth between corpus-level trends and specific textual instances that exemplify or embody those trends. That combination of approaches is often useful and rewarding, but our method here is different. We combine textual and social or paratextual traits, both of which we examine across a medium-sized corpus as a revealing case study. Our aim is to understand better the interplay between specifically medium- and large-scale textual and social developments during a particularly fast-changing period of Scandinavian (and modern) history. In this way, we seek to address Alan Liu’s decade-old call for a more culturally oriented Digital Humanities, as well as to continue the wider literary field’s long-standing investigation of literature as a historically conditioned, socially symbolic act.1

Our case study includes more than 800 Danish and Norwegian novels published between 1870 and 1899. The connected metadata about these volumes, manually curated on the basis of bibliographies, booksellers’ catalogs, and literary histories, contains biographical and bibliographical information on the novels and their authors (author’s gender, publisher, page length, sale price, etc.). These carefully collected social facets allow us to examine the uneven rises of modern Scandinavian literature across multiple intersectional aspects of its historical development.

2 | CONTEXT: “THE MODERN BREAKTHROUGH” AND ITS TENSIONS

The late nineteenth century is the time of the so-called “modern breakthrough,” in which Scandinavian societies underwent profound social and cultural transformations, and in which the literary field underwent similarly radical changes, with the formation of new aesthetic positions (realism and naturalism), with new authors stepping up from more diverse backgrounds (most importantly female writers, cf. Dahlerup, 1983; Jensen, 1993) and with an increasing literary output (not least consisting of novels). Danish historian Ole Thomsen defines the modernization of culture, politics, and daily life in the decades between 1870 and 1914 as “the great moult” (Thomsen, 1998, 10). In literary historiography, the common designation of the period, often demarcated by the decades of the 1870s, 1880s, and 1890s, is the “modern breakthrough,” named after a book title from the time by prominent critic and intellectual Georg Brandes (Det moderne gjennembruds Mænd [The men of the modern breakthrough], 1883). The fact that this label, at least in a Danish context, has become somewhat commonplace, across fields and domains, seems to testify to a defining role of literature in the era. Thus, the common and absolutely reasonable short version of the aspirations of Scandinavian literature of the modern breakthrough emphasize how it, in dialogue with contemporary European thought, represented a cultural reflection of societal transformations, echoing or emphatically engaging with them through artistic experiments (Ahlström, [1946] 1973). However, such notions of
literature’s newness and of its central role in society can of course be challenged and nuanced as well as qualified and quantified, something which this article tentatively aims to do.

For broad context, two megatrends and fundamental tensions between something old and something new shaping nineteenth-century literature and evidently present in and around the literature of the modern breakthrough can be highlighted. First, aesthetically, in the latter part of the nineteenth century decisive interventions were made in the larger, gradual movement from idealism to realism. From being detached from historical reality and demands of political and moral engagement, literature was designated to intervene in the reality of which it is part (see Bjerck Hagen, 2019). After the modern breakthrough, realism, understood as a noun rather than an adjective, as a historical, avant-garde movement became the norm in Scandinavia, Per Stounbjerg has argued (Stounbjerg, 2002, 14). Second, in terms of the sociology of literature, following from an expansion and diversification of the book market, the formative structures of a “literary field” in the sense that Pierre Bourdieu has accounted for in mid-nineteenth century France became apparent in Denmark-Norway with Copenhagen as a center: the split between a “commercial” or heteronomous pole and a “pure” or autonomous pole, in its purest form, populated by writers motivated by literary ends alone, turning their backs on the audience (Bourdieu, 1993; Speller, 2011, 79–101). These two roads, towards aesthetic heteronomy vis-à-vis social autonomy, crossed emphatically in Brandes’s famous 1871 introduction to his lecture series on Main Currents of Nineteenth-Century Literature. Brandes propagated a literature “that provokes debate” while positioning himself against the popular, but useless “literature of the moment,” flooding the book market and the daily press (Brandes, [1872] 2017, 700).

3 | DATA: A LOT OF NOVELS (BUT NOT ALL)

On a less abstract level, facts and numbers concerning late nineteenth-century literary production provide important context for an understanding of “the modern breakthrough,” which goes beyond Brandes’s somewhat exclusive and excluding concepts. The dataset underlying this investigation consists of all first editions of novels by Danish and Norwegian authors published in Denmark between 1870 and 1899, totaling 839 novels or approximately 55 million tokens.² Noteworthy is an explosion of reading matter, which to a large extent was fuelled by newspapers (Danish titles increased from 36 in 1847 to 156 in 1914, “the golden age” of the daily press in Denmark, cf. Sellinge & Thomsen, 1989, 48), but also by novels as indicated by the distribution of the 839 novels of our corpus (see Figure 1).

We see a substantial growth in the output of novels 1870–1900, from about 10 or 20 to more than 50 a year. Clearly, the consumption of novels scales up, mirroring and feeding the reading explosion. But the question is, what kinds of novels were produced and consumed?

A provisional explanation, derived from the critical conventions of literary historiography in which realism (including sub-movements of realism such as naturalism and impressionism) dominates the story, would be that new aesthetics is the driver behind this, since the novel as form favored realism (together with drama). However, since the novel had a double status not only as high art but also as popular culture,³ there is probably at least another, opposite, factor at play behind the raw output numbers, suggesting not only the heterogeneity, but also the social depth and distinctions of the texts. A glance at the titles and subtitles—for example, Klokken i Rosendalen. Original romantisk Fortælling (1876, The Bell in the Valley of Roses. Original Romance), Antonio eller Bjergenes Konge. Folke-Roman (1891, Antonio or, The King of Mountains. A Popular Novel), and so on—indicate that literary production in the period is packed with romances, historical novels, and other popular (sub)genres. A possible step towards nuancing conceptions of the modern breakthrough is to take into account—figuratively and literally—the popular literature surrounding and thereby, following Bourdieu’s theoretical framework, socially defining and distinguishing artistic literature. Among the 839 novels in our corpus only a small fraction plays a role in later literary historiography, whereby in a tacit way, compared to Brandes’s outspokenness towards popular literature, the bulk of literary production and of popular consumption is left out of the picture. The result of this is a most exclusive canon.
Moreover, the social and demographic properties of the producers can add significant nuances to our understanding of the literary period. The influx and diversification of novels and other reading matter were accompanied by radical changes in the population of writers, with new groups of authors stepping forward from diverse backgrounds and orientations: journalists and teachers (Kristensen, 1970 [1945]), but most noticeably women. Figure 2 shows the gender distribution of the authors of the novels over the 30-year period.

It indicates an increase in the number of books by women, though the fraction of total output by women remained relatively constant across the period. The ratio of female contributors to the corpus could potentially have been higher, if we had succeeded in identifying more of the authors writing anonymously or pseudonymously. Judging from the subset of such novels for which we have been able to attribute authorship by the help of external
sources, women were significantly more likely to hide their name on the title page than men (while 20% of the "named" novels in the corpus are by female authors, 27% of these anonymous and pseudonymous novels are authored by women). Nonetheless, the canon of novel-producing writers is almost exclusively male, with Jens Peter Jacobsen, Herman Bang, Holger Drachmann, and Henrik Pontoppidan as the dominant Danish quartet, and Henrik Ibsen, Bjørnstjerne Bjørnson, Alexander Kielland, Jonas Lie ("the Four Greats"), as well as Knut Hamsun as the Norwegian protagonists. These are the ones to whom separate chapters in literary histories—or street and square names for that matter—have been dedicated.

Our corpus opens up new possibilities of studying relationships between the few texts we know, which often correspond with the above-mentioned canon of contemporary novels by a limited group of male authors, and the many unfamiliar ones, whether trivial, for example the first Danish instances of Conan Doyle-esque detective fiction as recently mapped by Palle Schantz Lauridsen (Lauridsen, 2022), or radical, as exemplified by Otto Martin Møller’s Nina. En psykologisk Skildring (1883, Nina. A Psychological Portrait) rediscovered a few years ago by Heede (2019), which apparently is the world’s first novel on lesbian love.

A distant, socially informed reading of novels from 1870 to 1900 relying on paratextual and bibliographical evidence testifies to a high degree of heterogeneity in terms of both the aesthetics of the texts and of their authors’ sociology. The numbers suggest that the modern breakthrough, if nothing else, was the breakthrough of the novel in Denmark as the numbers in Figure 1 indicate. A further declaration of data is needed, however, before assessing aspects of the cultural width of the many novels. With different bibliographical data, comprising not only domestically produced but also translated novels, Franco Moretti has told a different quantitative story, which exposes a rise of the novel in Denmark in the 1820s (Moretti, 2003, Figure 4, p. 74). Focusing as we do on newly and locally produced literature—and thereby omitting not only translations but also reprints and, not least, serializations in newspapers and magazines—our corpus captures a particular aspect of literary history, one that emphasizes authorial production over readerly consumption. There are a lot of never-noticed or forgotten novels in there, but the corpus does not represent everything, as illustrated by the case of the once popular, pulp novelist, Ludvig Møller (1832–1904), with the pen name Louis de Moulin (a literal translation of his Danish name). Møller is represented with four novels in the corpus, for instance Jeanette Miron eller Fjorten Millioner (1877, Jeanette Miron or, Fourteen Millions), but from 1879 this "Danish Dumas" allegedly produced singlehandedly more than 400 novels for the feuilleton magazine Revuen (Hougaard, 1996, 10–11). In other words, like most literary datasets, our corpus does not record "what was actually read," to borrow a phrase from the book historian William St Clair (St Clair, 2004, 14). Correspondingly what we are looking at here, as in similar digital literary investigations which typically rely on datasets of first editions, is what Andrew Piper has called "writerly behavior" rather than readerly (Piper, 2017). You cannot have it all, and there are all kinds of sound pragmatic reasons behind this practice of data selection and corpus building. (Importantly popular texts such as newspaper serials are often beyond bibliographical control as well as difficult to digitize to the complex publication context constituted by newspapers.) Of course, you must not have it all either, but the ideological implications connected to the bibliographic criteria of selection must be acknowledged, not least when it comes to making quantitative claims about the social distribution of literature. With this in mind and considering the data, and not least the metadata, we have at our disposal, we want to test and discuss different strategies of explanatory modeling of the social depth of literary production and consumption in the period.

4 | BASIC EXPERIMENT: MODELING REALISM

4.1 | A theory of realism

As a test case trying to make sense of literary developments and social tensions within a large heterogeneous body of texts, we want to investigate the distribution of literary realism in the 30-year period spanned by our
corpus. The advancement of realistic prose is fundamental to traditional understanding of the period's literature, but we lack quantitative assessments of its temporal aspects (When did realist discourse kick in?) as well as its centrality (How widespread was realism?) and social distribution (Who were the producers of literary realism? And who were the consumers?). However, realism is a complex and elusive phenomenon, denoting both a universal and a historical form with ambitions of verisimilitude or mimesis as well as both methods and attitudes of authors confronting and shaping reality in their writing. In the following, tentatively, we want to emphasize certain formal traits of realistic prose, which can be summarized under the headline "manifold of details." This aspect of realism rings well with both late nineteenth-century aesthetics and later theorizations of the concept. Further, the richness of detail is a stylistic feature, which offers possibilities of counting and computing. In this heuristic definition realism becomes tangible in every sense of the word.

In a groundbreaking essay from 1921, the linguist Roman Jakobson encircled realism as an art form that aims at "conveying reality as closely as possible and strives for maximum verisimilitude" (Jakobson, [1921] 1987, 20). Interestingly, Göran Printz-Påhlson has pointed out how Jakobson, in one of his postscripts with specific respect to realist literature, clarified this definition by emphasizing "the densification of the narrative through images based on proximity" (Printz-Påhlson, 1989, 12). Such a conception of literary realism as an "obsession of detail, often trivial or detached" (p. 12) goes hand in hand with the self-understanding of realist aesthetics as formulated by its practitioners in the late nineteenth century. Two Scandinavian essays on realism testify to this. In 1882, in a newspaper manifesto "On realism," the Swede August Strindberg defined the phenomenon as the form in which "the artist tries to develop the intended impression, that is, to create an illusion, by performing the most important of the manifold of details of which the picture is composed" (Strindberg, [1882] 1913, 192; our translation). In 1879, Danish novelist and critic Herman Bang had concurred: "The manifoldness of the observed is too overwhelming" (Bang, [1879] 2001, 23). Famously, such a principle of redundancy noted and discussed by Strindberg and Bang was given a philosophical underpinning by Émile Zola in his essay "The experimental novel" (1893), in which "the objective reality of things" was stressed as a condition for life and literature (Zola, [1893] 1964, 34).

An example of realistic narrative in the style of Strindberg, Bang, and Zola, which illuminates the manifold and redundancy of detail and things, while also supplementing the male artistic gaze with a female one, is the opening of the Norwegian author Amalie Skram's short story "Madam Høiers Leiefolk" (1882, Madame Høier's Lodgers):

Below was the horse stable, and above a family dwelling, and the house lay in one of the simplest parts of the town. The glass in the windows was bottle-green, and the curtains of the window boxes black-spotted and coarse-grained with long unplugged wings. The blinds were rodless; they hung and dangled with their trapezoidal edges, and had never been properly rolled up; it was evident from the slanting fall which would not be remedied.

(Skram, [1882] 1892, 3; our translation)

Skram's setting of the scene displays what James Wood refers to as the “grammar of realism,” by having a surplus of both telling detail, including what Wood calls “tellingly untelling detail,” the stuff that “fills out a world, creates an atmosphere, and helps us to read a scene and situate it in a recognizable reality” (Wood, 2021, 133). In a discrete way the reader gets an impression of the social and psychological situation of the household, in which the story of the (poor) lodgers will unfold. Further, with the many adjectives and complex noun phrases Skram's text exemplifies Wood's observation of the "preference for concrete over abstract or generalized specification" in realistic prose (p. 133), echoed by Strindberg in his critical stance towards old-school idealism in the 1882 essay: “The writers of the previous era could make impressions through general indications" (Strindberg, 1913, 192; our translation). Both the (modern) degree of detail and the particularity in description are reflected by the many neologisms in Skram's prose. Two of the compound words in the cited passage, the hardly translatable "sortsjollede" (black-spotted) and "grovgrisne" (coarse-grained), are not present elsewhere in a corpus of more than 55 M words.
4.2 | Technical methods

But how do we model this “manifold of detail”? Indeed, how can we know which of the novels in the corpus are realist and which are not? Realism is not among the facets of the existing bibliographic metadata, and most of the books in the corpus are obscure. Even if we did possess a complete bibliography of realism in the period, we could not be sure that it was assembled on principles mirroring our own.

For these reasons, we elect to build a supervised computational model that we can use to predict the probability that any book in our corpus is an instance of realist fiction. We begin by assigning expert labels to 29 volumes (15 realist, 14 non-realist). These labels were preregistered, meaning that they were fixed before we ran any of our experiments and could not be changed in response to our later findings.

We use a simple logistic regression model combining token-based and linguistic features. Specifically, we measure the relative (length-normalized) frequencies of the 45 most informatively distributed words (as determined by ANOVA F value) between the two groups of labeled texts (realist and non-realist), as well as a set of 29 linguistic features that include the relative frequencies of basic parts of speech, the number of named entities mentioned, the length of the text (in words), and a few others. This mixed model outperforms one that uses either tokens or linguistic features alone. Indeed, it achieves 100% accuracy under cross validation on the labeled training data. The observed level of accuracy does not imply, of course, that the model is perfect. But it explains in part why we have not pursued more advanced, transformer-based methods, which are generally accepted as representing the state of the art in natural language processing. In addition to performing well, our model has the merit of being easy to introspect and to interpret, affordances of which we will make use of in our discussion below. Our model assigns to each of the 839 books in the corpus a predicted probability (between 0 and 1) that the book is an instance of realism. In combination with the collected bibliographic metadata, we use these predictions to observe changes in the relative level of realism over time and across categories such as author gender or nationality, publisher, and so on.

Separate from our predictive model, we construct four measures of “object density,” following the method of Tenen (2018), in order to test one version of the theory that realism is characterized by a profusion of object-level detail. These metrics are based on our ability to produce a dependency tree of each sentence in each book, and to select from it the concrete common nouns used as subjects and objects. From this set of nouns, we count those that DanNet (the Danish-language version of WordNet; see Pedersen et al., 2009) identifies as belonging to non-human, non-abstract classes (that is, as being concrete objects in the ordinary language sense).

We construct four such metrics, rather than one, in order to capture two different dimensions of the problem. First, we are interested (like Tenen) in the difference between the total number of such objects and the number of unique objects used in each book. Our intuition is that the number of unique concrete objects is more strongly associated with realist practice than is the total number of mentions of those objects, since one plausible interpretation of the “manifold of detail” is that it consists in the range of things that merit textual notice. So, we count both the total number of concrete object mentions and the number of unique concrete objects in each volume. We also select terms from DanNet in two different ways, owing to the imperfect fit between nineteenth-century usage and the contemporary sources that inform DanNet. In one version of the metric, we include only those common objects that are listed among selected, concrete synsets (sets of words that have similar meanings or other commonalities) in DanNet. In the other version, we include all objects unless they are listed among a contrasting set of human or abstract synsets. We find that the second of these methods is the more inclusive, though we defer discussion of this fact to our detailed analysis below.

5 | THE RISES OF REALISM

We have two main goals in our quantitative work. First, we want to trace the historical development and prominence of realist literature in Denmark and Norway, including the extent to which realism varied across social...
and commercial contexts. Second, we want to examine the features associated with realism in order to quantify and evaluate the long-standing critical consensus that realism is tied to the “manifold of detail” described in nineteenth-century manifestos and twentieth-century criticism.

We begin with a striking figure showing the rise of realism between 1870 and 1899 (Figure 3). At the beginning of the period, between 1870 and 1875, about 20% of volumes (or fewer) were realist. This fraction rises rapidly and steadily over the next two decades. By 1880, more than half of newly published books in the corpus were realist. By 1890, the share of realism was well above 80%.

The temporal plot matches literary-historical expectations perfectly. As expected, we have a steady rise of realism across the 30-year period with the early 1880s as breakthrough years, around which Bang and Brandes published their literary manifestos (Realisme og Realister, 1879, vis-à-vis Det moderne Gjennembruds Mænd, 1883). Arguably, the plot can be said to exceed or exaggerate literary developments, when we consider the fraction sizes.

To begin with, coming from traditional literary historiography, where the realism of the modern breakthrough is treated as an avant-garde phenomenon and defined on the basis of a few, prolific authors, a realist fraction of 20% is a lot, and the same can be said with the climax of 80% in the 1890s. The mismatch in proportions between our model and the implicit ones in literary histories can be explained both by the fact that we are dealing with much more data in our investigation, including novels by female authors, often barred in historical tradition, and, of course, by the fact that the concept of realism in our model is broader than the one implied in historiography. It could be speculated that our concept engulfs many realisms, both avant-garde forms, including sub-branches such as naturalism and impressionism, and more traditional, in a Danish/Germanic context, so-called “poetical realism.”

Regardless of how this new history of the historical distribution of realism relates to historiography, it is qualified by the fact that the premises and calculations behind it are both straightforward and transparent. Of course,
Figure 3 represents a summary view, which smooths over most of the detail found in individual books and groups of books. The story must be further faceted.

To begin with, to get a better sense of how the model performs and, more importantly, to gain a better sense of the poles of realist and non-realist narrative discourse in the period, we can examine the volumes that our model predicts to be most and least realist, as measured by their probability of belonging to the realist class. These books are listed in Table 1 (realist volumes) and Table 2 (non-realist volumes) below.

Of course, a lot can be said of each of the 20 novels over- or underperforming as instances of realism. Standing out as probably the most familiar and the only high-canonical text on the list, is Henrik Pontoppidan's first individually published novel *Sandinge Menighed* (1883, Sandinge Conggregation), which depicts the social and human misery both in the country and in the city, forming a point of departure of a momentous realist oeuvre (awarded the Nobel Prize in 1917). Pontoppidan's novel is appropriately placed as a high-scorer in the realist division. But what really gives merit to the model are the general trends indicated by the contrasting lists of books, given away by paratextual and contextual clues. In terms of year of publication, strikingly, all the non-realist novels are published

### TABLE 1 Realist volumes. Books with the highest predicted probability of realism.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Author</th>
<th>Title</th>
<th>Subtitle</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Vedel, Valdemar</td>
<td>Stavnsbaand</td>
<td>Fortælling</td>
<td>1888</td>
</tr>
<tr>
<td>2</td>
<td>Pontoppidan, Henrik</td>
<td>Sandinge Menighed</td>
<td>En Fortælling</td>
<td>1883</td>
</tr>
<tr>
<td>3</td>
<td>Møller, Niels</td>
<td>Hændelser</td>
<td></td>
<td>1890</td>
</tr>
<tr>
<td>4</td>
<td>Bon, Fridtjof</td>
<td>Stoltenberg og andre Svende</td>
<td></td>
<td>1898</td>
</tr>
<tr>
<td>5</td>
<td>Alsted, Peter</td>
<td>Ly</td>
<td></td>
<td>1897</td>
</tr>
<tr>
<td>6</td>
<td>Brummer, Therese</td>
<td>En Kamp</td>
<td>Familieinteriør</td>
<td>1893</td>
</tr>
<tr>
<td>7</td>
<td>Petersen, Lauritz</td>
<td>Mellem Klitter</td>
<td>Roman</td>
<td>1894</td>
</tr>
<tr>
<td>8</td>
<td>Winterhjelm, Kristian</td>
<td>Naturalisterne</td>
<td>En Sommerskitse fra Stockholms Skjærgaard</td>
<td>1886</td>
</tr>
<tr>
<td>9</td>
<td>Møller, Niels</td>
<td>Koglerier</td>
<td></td>
<td>1895</td>
</tr>
<tr>
<td>10</td>
<td>Nøllsen, Jappe</td>
<td>Nemesis</td>
<td>Fortælling</td>
<td>1896</td>
</tr>
</tbody>
</table>

### TABLE 2 Non-realist volumes. Books with the lowest predicted probability of realism.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Author</th>
<th>Title</th>
<th>Subtitle</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>839</td>
<td>Gundersen, W. K.</td>
<td>En Politimands Erindringer</td>
<td>Original historisk Roman med Portræer</td>
<td>1870</td>
</tr>
<tr>
<td>838</td>
<td>Sørensen, Carl</td>
<td>Alt for Fædrelandet</td>
<td>Original romantisk Fortælling</td>
<td>1878</td>
</tr>
<tr>
<td>837</td>
<td>Kofoed-Hansen, H. P.</td>
<td>Livsømner</td>
<td>En Fortælling</td>
<td>1875</td>
</tr>
<tr>
<td>836</td>
<td>Albertius, Claudius</td>
<td>Klokken i Rosendalen</td>
<td>Original romantisk Fortælling</td>
<td>1876</td>
</tr>
<tr>
<td>835</td>
<td>Vendé, James</td>
<td>Leonie eller En Nat i Merne (grundet paa vir...)</td>
<td>Original romantisk Fortælling</td>
<td>1875</td>
</tr>
<tr>
<td>834</td>
<td>Albertius, Claudius</td>
<td>Bjergmandens Ring 2</td>
<td>Original romantisk Fortælling</td>
<td>1874</td>
</tr>
<tr>
<td>833</td>
<td>Sørensen, Carl</td>
<td>Kjærleked ved Hoffet</td>
<td>Fædrelandshistorisk Skildring fra det atende...</td>
<td>1875</td>
</tr>
<tr>
<td>832</td>
<td>[anonymous]</td>
<td>Taterkongen</td>
<td>Original Fortælling</td>
<td>1876</td>
</tr>
<tr>
<td>831</td>
<td>[anonymous]</td>
<td>Kjærlighed, Hævn og Frelse</td>
<td>En original Fortælling</td>
<td>1872</td>
</tr>
<tr>
<td>830</td>
<td>[anonymous]</td>
<td>Capercapitaine</td>
<td>Englænderne i Kjøbenhavn 1801 og Kjøbenhavns B...</td>
<td>1871</td>
</tr>
</tbody>
</table>
in the 1870s, whereas none of the realist ones are published before 1883. Further, if we look at the producers of the novels, the non-realist authors (half of which are anonymous or published under a pseudonym: Claudius Albertius) are completely unknown and unmentioned in literary histories. The group of realist authors, on the other hand, does not form a canonical armada, but they were all established writers and literati with artistic ambitions and solid literary networks, one of them, an early writer of worker’s literature, Lauritz Pedersen, perhaps on the brink of a renaissance. Finally, the paratexual categories of title and subtitle give a clear indication of a genre system among the novels with equally clear cultural distinctions. In terms of generic affiliation, all of the predicted non-realist novels are clearly marked—or marketed, rather. If we think of Claude Duchet’s definition of the book title as “a coded message in a market situation” (cited after Moretti, 2009, 134), these 10 titles are just as much advertisements and pieces of pedagogical consumer advice as artistic signs. These non-realist novels belong to the realm of popular literature or what is sometimes called “genre fiction.” They are either historical novels or romantiske fortællinger (romantic stories), in other words, novels which in their narrative setting are displaced from the contemporary reality either in time or location. On the other hand, the predicted realist novels reveal themselves as present in their temporal and geographical setting. Otherwise their paratextual signals are mixed and heterogeneous, indicating artistic individuality and, perhaps, realist particularity.

Despite the temporal consistency of our observed rise of realism, there exist important differences in subsets of the corpus. Four of these are shown in Figure 4, which groups volumes by author nationality, publisher, typeface, and length (above or below the length of the global median volume in the corpus). The difference between the groups across years within each of these facets is large and statistically significant (Hedges’s $g > 0.7$, $p < 0.001$).

The faceted views of the distribution of realism over time show the importance of social and material contexts. A strict literary perspective must be complemented with considerations of the sociology of literature and of the literary marketplace in order to grasp and handle the driving factors behind the development. All the four decisive

![Figure 4](https://onlinelibrary.wiley.com/doi/10.1111/oli.12396)

**FIGURE 4** The rises of realism. Average annual predicted realism probability by four facets. Clockwise from upper left: length, typeface, publisher, and author nationality.
contextual factors above deserve a careful and lengthy examination, which is out of this article’s scope, so we limit ourselves to highlighting the most likely explanations behind the clear divisions in the data. If we look at the material factors to begin with, the fact that realism correlates with short book length could probably be explained both aesthetically and sociologically: whereas dealing with detail-dense realist literature took time and effort from both author and reader, putting a natural limit to the extent of the book, the actual costs of producing a book had been significantly lowered due to the industrialization of paper production and the printing and binding of books, whereby a long book no longer signaled exclusivity or cultural capital in itself (Bjerring-Hansen, 2015, 167), rather perhaps the opposite: popularity and cultural heteronomy. Further, when it comes to the question of typeface it is by no means surprising that books set in Roman type were more influenced by aesthetic advancements than their counterparts in Gothic type. Roman type was a clear signal of modernity and internationality, and the shift towards it taking place in this period was commanded by the leading publisher Gyldendal and liberal newspapers (see Bjerring-Hansen et al., 2022). Speaking of publishers, as expected, four imprints (Gyldendal, Schubothe, Philipsen, and Schou) were the main suppliers of realism. They appear on the title page of almost all canonized literature from the time (Jacobsen, Bang, Lie, etc.), and scaling things up radically does not change their position as central actors and arbiters of taste in the story. Finally, and much more tricky to interpret, is the factor of author nationality. Why is the Norwegian portion of novels much more influenced by realism than the Danish? We know the modern (realist) qualities of the most prominent Norwegian writers of the modern breakthrough as well as their love-hate relationship with Copenhagen, which constituted the cultural center of the region or, in the words of Ibsen, the “common literary capital” (cited after Fulses & Rem, 2016, 454). With 133 novels by 37 different Norwegian authors, however, our data goes beyond the more familiar cases. We can only speculate on the causality at play here, whether, for instance, Norwegian authors, major as well as minor, with realist inclination opted for Copenhagen (in order to reach a much larger audience or a higher degree of understanding for their work), or whether the Copenhagen gatekeepers tended to prefer realist writings, perhaps even a Norwegian-flavored realism? Again, a result like this calls for more critical attention, including both contextual scrutiny and close readings.

Several facets that we might have expected to show important differences in the distribution of realism do not do so. These include author gender, book price, and the presence or absence of illustrations in the text. Particularly the non-difference concerning gender as a facet is interesting and perhaps also surprising. An author like Amalie Skram, whose books all rank pretty highly in predicted realism, was framed rather negatively in the contemporary reception. Her style was too raw and direct, too “manly.” She was accused of copying Zola, whom she had never read (Garton, 2007, 311–328). This might lead one to believe that she was, in a statistical sense, an outlier, but the bias is no greater among women writers than men. In our model, realism was just as much the business of women as it was the business of men.

The four facets that do show important differences between groups are unlikely to be independent from one another. Publishers choose typefaces for the books they publish; authors from different nations have different access to the Danish literary market; readers’ collective taste for longer or shorter books changes over time, and so on. While we are not confident that we have sufficient historical and social data to build a full causal model to disentangle these factors, we do seek ways of measuring their relative influence.

One approach—which we adopt—is to build a multiple linear regression model that predicts the annual average output of our realism model on the basis of the annual mean level of each metadata facet, as well as the year itself. Recall that none of these paratextual facets is an input to our original model of realism. We are, in effect, treating the average predicted level of realism in our corpus in each year—which was a function of strictly textual features—as an independent observation, which we then attempt to model as a (linear) function of paratextual features, in order to understand better which of the paratextural features are more significant, controlling for the others. When we build this model, we find, first, that it performs remarkably well, having a coefficient of determination ($r^2$) of 0.947 (that is, about 95% of the variation in annual predicted realism can be captured through variation in our paratextual inputs). Only three of the input facets have large coefficients: typeface (Roman type is strongly predictive of realism), year of publication (later books are more likely to be realist than are earlier books),
and author nationality (years with more Norwegian authors generally contain more realistic texts). The other facets (author gender, text length, text price, and publisher) are not strongly associated with predicted realism. We emphasize, however, that this model is not causal; Roman type does not cause realism. But it does mean that there is little additional predictive information in, for example, the identity of a book’s publisher after we control for other paratextual facets that the publisher may control. If you wanted to know whether a book from our corpus was likely to be realist, without consulting the text itself, you would do best to ask about its typeface and date of publication, and about the nationality of its author.

6 | THE TEXTURE OF REALISM

The second part of our investigation concerns the properties of realism and the critical tradition surrounding them. As we noted in the section on technical methods, our text-based predictive model uses two types of features: word counts and linguistic properties. We can reason about what we expect to find among these features on the basis of widely studied realistic texts from the late nineteenth century, on poetological writings from the time, and on later critical definitions and readings. We hypothesized that the stylistic devices and linguistic features that most effectively distinguish realist from non-realist texts include: (1) a preference for uncommon over common words; (2) a preference for concrete over abstract nouns; (3) an abundance of descriptive adjectives and adverbs; and, finally, (4) an accumulation of nouns and noun phrases at the expense of verbs (i.e. what in German linguistics is called Nominalstil as opposed to Verbalstil), based on Sven Møller Kristensen’s assertion that “throughout the nineteenth century, there is a tendency towards the noun-based style [substantivisk stil]” (Møller Kristensen, 1955, 58).

We begin by examining the word frequency features in our model. Recall that these features are not ones we selected in advance. They were determined by the model itself to differentiate most efficiently the labeled realist and non-realist texts in our training set. Our task, then, is to characterize these features, and to evaluate them against critical expectations. To reduce the risk of telling a just-so story, we compared expert predictions of the category association of each of the 45 word features against the empirically observed category of each term. For example, the word bemærket ("noticed") is strongly associated with either realism or non-realism in our training data. But which one? (Non-realism, it turns out.) Without knowledge of the results, our expert correctly predicted the class of 42 of the 45 features (an error rate of just 6.7%).

Confident that the word features can thus be reconciled with critical expectation, we note that adverbs (altid = always, fint = fine, naturligvis = of course) and adjectives (i.e. nervøst = nervous and stive = stiff) are associated with realist novels, whereas abstract nouns (Vilje = will, Leilighed = occasion) as well as all verbs were determined to be non-realist features. It is also the case that the majority of the word features occur more frequently in non-realist novels than in realist texts (35 of the 45 features, about 78%). This makes sense—if we credit the professed realist ambitions towards distinctive description. That ambition should have produced a more dispersed or wide-ranging vocabulary in realist texts, leaving the most concentrated terms to occur frequently in non-realist volumes.

On the whole, we find that the observed word features accord well with existing critical models of realism. The same is broadly true of the most distinctive linguistic features, which likewise fit a critical story that privileges the role of material detail. Top features, as measured by Shapley values, include verb and named location frequencies, determiners, multi-word noun phrases, and text length (all strongly associated with non-realism) and adjectives, adverbs, and coordinating conjunctions (strongly associated with realism). This makes broad sense if we think of non-realism as more often concerned with action, in contrast with realism’s preference for detailed portraiture. In non-realist books, actions happen to and with generic objects in locations of some type. Non-realist books are longer, on average, to accommodate their correspondingly more extensive plots.
Realist books, by contrast, describe their objects in greater detail, eschewing (relatively speaking) events in favor of dwelling on objects.

We note, however, an unexpected result, namely that realist texts are not enriched in nouns or noun phrases. If realism is characterized by portraiture and by the attempt to represent a "manifold of detail," is it not surprising to find realist texts no richer in nouns than their non-realist counterparts? To understand this result better, we turn to our object density measures. There are two things to keep in mind as we do so. First, the object density metrics are not among the inputs to our model of realism. This does not mean that the model is entirely unresponsive to features related to object density, since it includes correlates like noun and named-entity frequencies, but we use the model to divide the corpus into realist and non-realist subsets, and only then examine an independent feature (object density) that we expect to be associated with realism. Second, the object density metrics are designed to exclude human beings, abstract nouns, and all proper nouns, as well as any nouns that are not used in context as subjects or objects. They are, in short, meant to capture the material or commodity density of the text, a different thing than noun density. A comparison of the object density metrics is summarized in Table 3.

In Table 3, “exclusion” and “inclusion” refer to the process by which we match textual terms to their categories in DanNet. In the inclusion case, we count terms that positively match DanNet lists of concrete object types, rejecting common nouns that do not occur on those lists. In the exclusion case, we exclude undesired nouns that match a complementary set of non-concrete (human and abstract) categories, counting those that remain. Differences arise in part from polysemy, but also (and more often) because DanNet was not built to accommodate the peculiarities of nineteenth-century literary text digitized via optical character recognition. The question, in effect, is how to treat words that our dependency parser and part of speech tagger identify as nouns of the relevant types, but that do not occur in DanNet. The exclusion-based count allows them, while the inclusion-based metric discards them. The “unique” versions of these metrics count the number of distinct objects that occur one or more times in a given text, rather than the total number of such occurrences.

Comparing the four object density measures suggests how it is possible for nouns to be fairly evenly distributed between realist and non-realist texts, while still maintaining a significant role for a profusion of material objects as a defining trait of realism. The non-unique versions of the metrics—which count every mention of a relevant object—show little difference between the two groups. The exclusion-based version, which errs on the side of including human and abstract nouns when they do not exactly match DanNet’s lists, most closely resembles a raw noun frequency count and is, indeed, the only version of the metric that matches noun-frequency’s association with non-realist volumes. The unique versions of the metric focus instead on the range of objects used (rather than the number of times that objects are mentioned). In these cases, there is a medium to quite large effect size in favor of higher values in realist texts.

We see the strongest effect in the case of the inclusion-based count of distinct objects. This is as we might have expected. The inclusion-based metric is the most restrictively focused on concrete nouns, and its unique version captures the number of distinct objects that rise to the level of narrative notice, rather than the total number of times objects are mentioned. Books that score highly in this metric are those that devote the largest fraction of their pages to detailed descriptions of inanimate objects.

**Table 3** Mean object density metrics (occurrences per 100,000 words) by realism class with effect size and significance level.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Realist</th>
<th>Non-realist</th>
<th>g</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exclusion</td>
<td>2,221</td>
<td>2,344</td>
<td>0.32</td>
<td>0.0002</td>
</tr>
<tr>
<td>Exclusion, unique</td>
<td>1,512</td>
<td>1,359</td>
<td>0.46</td>
<td>0.0002</td>
</tr>
<tr>
<td>Inclusion</td>
<td>769</td>
<td>762</td>
<td>0.05</td>
<td>0.5090</td>
</tr>
<tr>
<td>Inclusion, unique</td>
<td>387</td>
<td>296</td>
<td>0.96</td>
<td>0.0002</td>
</tr>
</tbody>
</table>
of their narrative attention to the material details of the physical world. They are, in short, good candidates for volumes that attempt to capture the manifold of detail of the world they depict.

Focusing, then, on the inclusion-unique object density metric, we can return to the paratextual socio-economic facets that we examined above (see Figure 4 and associated discussion). We find similar results. Object density in general rises smoothly and significantly over time, from fewer than 300 unique objects per length-normalized volume in the early 1870s to around 400 by the 1890s. Part of the explanation is decreasing volume length: late-period books, although generally shorter, were packing more objects into less narrative space. Books set in Roman type were much more object-dense than were their Fraktur counterparts, as were cheaper volumes and male-authored books. When we control for covariability via a multiple linear regression model, as we did in the case of predicted realism, we find that year of publication and text length are the most strongly predictive factors, followed by price and author gender. Recalling that neither price nor author gender was strongly associated with predicted realism overall, we suggest that critics might profitably return to their own internal models of the co-involvement of realism, gender, and the literary market. Have we, collectively, overemphasized the role of object-level detail in identifying canonically realist texts? Might we formulate a more broadly inclusive conception of realism that responds to the breadth of its rise across social facets through the late nineteenth century?

7 | CONCLUSIONS

Our investigation of the rise and impact of realism in Danish and Norwegian novels from 1870 to 1899 shows that Stounbjerg is right in his assertion that realism became the “norm” in Scandinavian literature after 1870, perhaps more right than he would have expected. According to our model, around 1870 some 20% of the novels were predicted to be realist. By 1880, it was more than half, while by 1890 the share of realism was well above 80%. It can be argued that our model is too inclusive and that it deprives realism of its avant-garde or elite character, but on the other hand it can also be stated that traditionally, based on very little empirical material, we have operated with a concept that is far too exclusive. Either way, it is not a question of competition or of either/or. The dialogue between analog and digital approaches, between hand picking and machine picking, is not made less interesting by such divergences. On the contrary, we believe.

We have concentrated on major trends and their components, rather than individual authors and specific varieties of realism. But others can access our code and our results on their own and, just as an example, examine how the novels of more renowned authors relate to realism, in general or in terms of specific aspects (e.g. the frequency of adjectives or of unique nouns). In this context, we could focus, among other things, on different currents and individual stylistic attitudes within the complex phenomenon of historical realism, for example, J. P. Jacobsen’s heavy use of adjectives versus the adjectival aversion in Herman Bang’s prose (the latter distancing himself from the realism of the 1870s in the name, paradoxically, of realism; eventually he would label his dialogue-heavy style as “impressionism”). Rather than this type of nuance, we wanted to bring contextual factors into play, to facet the analyses of a literary phenomenon using biographical and bibliographical information. In this context, interesting new patterns emerge, which can partly be understood as cultural “distinctions” in Bourdieu’s sense (the imprint, the length, and the typography of the books are indicators of their degree of realism), partly reveal complex cultural fractures (Norwegian authors were more influenced by realism than their Danish counterparts, while women did not stand behind men, despite objections from contemporary criticism). 16

We recall that our research had three, interlocking critical and methodological aims. We wanted to use textual features to model realism in a large corpus of Danish-language novels. We wanted to compare the features driving that model to the ones that were important to the historical and critical development of realism as a literary project. And we wanted to use the results of our model, assuming it performed well, to study the interplay between realism and social history.
Have we succeeded? That is not for us to judge. Our model is accurate within the limits of the training data we have generated and of our ability to evaluate its predictions. The model also appears to seize on many of the same features that critics have argued distinguish realism from earlier and less avant-garde literary forms. This convergence is by no means a given. It is entirely possible for a computational model to perform well while basing its decisions on features that human readers would not use to distinguish one text from another (Allison et al., 2011).

We hope that we have illustrated a process by which critical knowledge can be examined with reference to computational models, and in which computational models, in turn, are evaluated against critical knowledge, as Andrew Piper and others have both argued and illustrated at some length (Piper, 2015). We want also to emphasize the difficulties of studying the large-scale distinctions in realist writing across socially conditioned groups in the absence of computational modeling. It is something of a commonplace in the critical reception of computational literary work to emphasize the opposite aspect of this relationship: that text analysis may leave out historically pertinent information that cannot be extracted from the text alone. We agree. But, in this case, we have used a text-based computational model to help us make use of the socio-historical information that we do possess about the books in our corpus. Only with the help of our model could we distinguish realism from non-realism in more than 800 novels, and only with this ability could we study the conjunction of these particular socio-textual formations.

Finally, we are pleased to have been able to provide a worked example of such interwoven computational-critical work at large scale in a non-English target language and addressing a historical period very different from our own. While we do not share the view that Digital Humanities and cultural analytics research is uniquely Anglophone-centered, it is certainly true that English is well represented in literary studies and natural language processing alike. But the technical methods we have used are generalizable across many languages and, where they rely on language-specific resources (such as trained language models and DanNet), those resources exist for many (though certainly not all) languages.

In the end, our findings suggest that realism was more prominent and more widely distributed in Danish-language novels of the late nineteenth century than the critical tradition has usually acknowledged. Women appear to have written realist fiction not only at rates similar to men, but at times and with features that are difficult to distinguish from their male counterparts. If critics have found it difficult to perceive these broadly distributed rises of realism, it may be in part because they have worked with an avant-garde-ish concept of realism that was explicitly restrictive and that was highly focused on objects to the exclusion of other features even within the books that did merit their attention. The view of the modern breakthrough that we have produced is necessarily incomplete, but we hope that it can serve as an occasion to return to the socio-literary dynamics of the period with new tools and new knowledge.

FUNDING INFORMATION

The research underlying this article was supported by the Carlsberg Foundation.

ORCID

Jens Bjerring-Hansen  https://orcid.org/0000-0001-5786-8300
Matthew Wilkens  https://orcid.org/0000-0001-6749-9318

ENDNOTES

1 On the relation between computation and cultural studies, see Liu (2012). For productive examples in practice, there are dozens in literary studies proper and thousands in the broader field of computational social science. See especially Bode (2018), Tatlock et al. (2018), and Underwood et al. (2018).

2 The corpus we are analyzing, version 0.5 consisting of 839 volumes, was curated, compiled, and processed by Jens Bjerring-Hansen, Philip Diderichsen, and Dorte Haltrup Hansen with the help of Nanna Emilie Dam Jørgensen (University of Copenhagen) on the basis of PDF scans and OCR provided by the Digitization Unit at the Royal Library, Copenhagen. The compilation of novels has been done on the basis of Dansk Bogfortegnelse, recording all publications by Danish publishers since 1830, including “Danish-Norwegian” language (dansknorsk) books by Norwegian authors, all of which are available at the Royal Library through legal deposit. In terms of data selection,
special attention has been paid to discard non-novels from the corpus (especially short story collections which are placed in the same section, "Fortællinger" (stories), as novels in Dansk Bogfortegnelse). In terms of text processing, special care has been taken to improve OCR accuracy in the subset of the corpus set in Fraktur. In the forthcoming version 1.0 of the corpus, the few remaining short story collections will be excluded, while a dozen texts with flaws (e.g. missing pages) will be included. For details about the corpus building and post-correction procedures, see Bjerring-Hansen et al. (2022).

2 This is a main argument in George Levine's How to Read the Victorian Novel (Levine, 2007).

4 We note that this result differs from previous work on nineteenth-century British fiction, which found that women did not use anonymity or gender-concealing names (including initialisms) at rates higher than men (see Evans & Wilkens, 2023; Jockers & Mimno, 2013).

5 Moretti's Danish data comes from Munch-Petersen (1978, 978–980).

6 See also the book historian Katherine Bode’s criticism of lack of bibliographical awareness in the curation of literary datasets, A World of Fiction: Digital Collections and the Future of Literary History (Bode, 2018).

7 Following the question (and the pragmatic answers to it) why not all novels are included, one could of course also argue that other kinds of texts than novels should be included in the corpus underlying this study. This is particularly relevant for drama, which was undoubtedly an important vehicle for the epoch’s turn towards reality and realistic endeavors. In an alternate study—one that is not currently possible given bibliographic constraints—the short prose genres of the time might have been included ("Noveller," "Noveletter," "Skisser," "Billeder," "Stemninger," "Krøniker," etc.). On the impact of short prose on the innovation of the literary conventions of the modern breakthrough, see Jørgensen (1991) and Auring and Svendsen (1999).

8 The 1882 short story was included as an appetizer to Skram’s marriage novel Forraadt (Skram, 1892, Betrayed) and, thus, it is part of the corpus, exemplifying the mentioned challenges of selecting and curating a homogenous corpus of literary texts.

9 For full details of the features and methods used, see the online code appendix to this article at https://github.com/wilkens/memo-realism.

10 For full lists of synsets in each group, see the online code appendix to this article.

11 These are the volumes about which the model is most confident. Each of the listed realist volumes has a predicted realism probability >99.8%. Listed non-realist volumes have a maximum realism probability <0.04%.

12 In 2018, Gadens Roman (1896, Novel of the Street) by Lauritz Petersen was republished by Gladiator in the revisionist series of classics from the decades around 1900, Sandalserien. Petersen’s novel ranks high in the list of realistic books, too (no. 26).

13 The first of these statistics is a measure of effect size. Effect size metrics quantify the relative (as opposed to the absolute) magnitude of the difference between two groups. They are generally computed as a variation on the difference between the means of the groups in question, divided by the pooled standard deviation of the data across those groups. By convention—though it is nothing more than convention—effect sizes above 0.5 are considered moderate, while those above 0.8 are considered large. Hedges’s g, used here, is similar to the more familiar Cohen’s d measure of effect size, with a small adjustment applied for better performance in small samples. Our corpus is sufficiently large that g is effectively equal to d in all cases; p values of statistical significance are estimated via permutation. In small samples, effect sizes generally correlate closely with tests of statistical significance (large effects align with high significance). In large samples, small effect sizes may produce highly significant results. For this reason, effect sizes are often preferred to p values as estimates of the practical importance of a difference between groups.

14 By extension, it is not surprising that many of the books that our model predicts to be non-realist are not only long in terms of pages, but also part of a series, e.g. Bjermandens Dotter 1 and 2 (1874, The Mountain Man’s Daughter) or Klokken i Rosendalen 1 and 2 (1876, The Bell in the Rose Valley).

15 Shapley values measure the cooperative game-theoretical contribution of each input feature to model output. In the present case, the rank order of their contributions tracks relatively closely with that of the directly inspectable coefficients of our logistic model. None of our conclusions would be reversed if we instead measured feature importance via direct inspection of the coefficients. For a detailed explanation of the use of Shapley values to interpret machine learning models, see Lundberg and Lee (2017).

16 The cultural distinctions between Danish and Norwegian literature that the study reveals could of course be interesting to extend to a genuinely Scandinavian perspective by also including Swedish novels from the period. However, the language barrier—between Danish and Danish-Norwegian on the one hand and Swedish on the other—complicates things and calls for experimental multilingual NLP solutions.


AUTHOR BIOGRAPHIES

Jens Bjerring-Hansen (jbh@hum.ku.dk) is associate professor of Scandinavian Literature at the Department of Nordic Studies and Linguistics at the University of Copenhagen. He specializes in eighteenth- and nineteenth-century literature with a particular interest in the sociology of literature and quantitative literary history.

Matthew Wilkens (wilkens@cornell.edu) is associate professor of Information Science at Cornell University, New York. He studies computational methods for humanities text analysis with special interest in large corpora and broad historical range.

How to cite this article: Bjerring-Hansen, J., & Wilkens, M. (2023). Deep distant reading: The rise of realism in Scandinavian literature as a case study. Orbis Litterarum, 00, 1–18. https://doi.org/10.1111/oli.12396