Digital Studies /
Le champ numérique

Primorac, Antonija. 2023. "Distant Reading Two Decades On: Reflections on the Digital Turn in the Study of Literature." *Digital Studies/Le champ numérique* 13(1): 1–24. DOI: https://doi.org/10.16995/dscn.8855.

Open Library of Humanities

Distant Reading Two Decades On: Reflections on the Digital Turn in the Study of Literature

Antonija Primorac, Faculty of Humanities and Social Sciences, University of Rijeka, HR, antonija.primorac@uniri.hr
Rosario Arias, Universidad de Málaga, ES, rarias@uma.es
Roxana Patras, "Alexandru Ioan Cuza" University of Iași, RO, roxana.patras@yahoo.ro
Eva Eglāja-Kristsone, Institute of Literature, Folklore and Art/University of Latvia, LV, eva.eglaja@lulfmi.lv
Karina van Dalen-Oskam, Huygens Institute/University of Amsterdam, NL, karina.van.dalen@huygens.knaw.nl
Berenike Herrmann, University of Bielefeld, DE, berenike.herrmann@uni-bielefeld.de
Christof Schöch, University of Trier, DE, schoech@uni-trier.de
Pieter François, University of Oxford/The Alan Turing Institute, UK, pieter.francois@stb.ox.ac.uk

This article examines the ways in which distant reading, as a facet of the digital turn in the humanities, has affected the study of literature, with particular attention to the ways the digital turn has impacted the examination of authorship, genre, and style. In the process, it reflects on the ways in which distant reading developed both as a concept in the history of world literature and as a methodological approach that contributed to the evolution of computer-assisted study of literature.

Cet article examine les façons dont la lecture à distance, en tant que facette du virage numérique dans les sciences humaines, a affecté l'étude de la littérature, avec une attention particulière aux façons dont le virage numérique a influencé l'examen de la paternité, le genre et le style. Dans le processus, il réfléchit sur les façons dont la lecture à distance a développé à la fois comme un concept dans l'histoire de la littérature mondiale et comme une approche méthodologique qui a contribué à l'évolution de l'étude assistée par ordinateur de la littérature.

1. Introduction

Since its introduction as a term in 2000 by Franco Moretti, "distant reading" has gone through many transformations both as a concept and as a method. These changes went hand in hand with the rapid impact of the digital turn in the humanities on the one hand and the growing application of digital tools to the study of literature on the other. In the last decade alone, distant reading provoked numerous theoretical debates about the meaning, purpose, and practice of literary analysis and generated the development of new approaches in the computer-assisted study of literature, from macroanalysis (Jockers 2013) to data-rich literary history (Bode 2017). Initially meant as a provocation within the field of comparative literature, distant reading gradually transformed from a call to re-conceptualize Goethe's idea of *Weltliteratur* beyond the comparative study of several, usually European, canons by examining instead all world literatures as an uneven and unequal but nevertheless single interconnected system through a combination of Wallerstein's world systems theory, evolutionary theory, Marxism, and formalism (Moretti 2013), to become synonymous with the study of literature founded on large-scale corpora analysis, exemplified by the Stanford Literary Lab pamphlets, and now sometimes also called computational literary studies. "Large-scale corpus" is here understood in the light of Allen B. Riddell's definition of a large collection as "any collection of texts … if it contains more texts than a single researcher would be expected to digest in a year's worth of dedicated reading" (Riddell 2014, 91). This article examines and reflects on the ways in which distant reading, as a facet of the digital turn in the humanities, has affected the study of literature, with particular attention to the ways the digital turn and the shift in scale has impacted the concepts of genre, authorship, and style.

In an article reassessing distant reading twenty years on, it seems appropriate to broaden the perspective on distant reading's nominal opposite—the concept of "close reading" itself. This "attentive inspection of the verbal texture of poems, novels, and plays" (Freedman 2015), became common practice at universities in the mid-twentieth century under the principles conceived and defended by I. A. Richards and William Empson. Richards's *Principles of Literary Criticism* (Richards [1924] 2017) and *Practical Criticism* (Richards [1929] 1978) were key in developing the method of "close reading," which was later taken up and disseminated by the "New Critics" on the other side of the Atlantic. Cleanth Brooks and John Crowe Ransom, for example, considered Richards as a figure of reference for New Criticism, which led critics to conflate Richards's tenets with those of the New Critics. Interestingly, Joseph North has recently pointed out the divergent directions that Richards and the New Critics followed in their methodologies, which remain largely unnoticed by current scholarship today. In his view, Richards did sustain a more external approach to the text, “directed towards an advanced utilitarian model of aesthetic and practical education” (North 2013, 142). It is true that Richards did not include any analysis of socio-political context, and that he only referred to extracts and short lyric poems in *Practical Criticism*, which constituted the basis of the “close reading” practice. However, he was aiming at considering the aesthetic as instrumental in exploring the relationships between literary works and their reception: “[w]e gain a much more intimate understanding both of the poem and of the opinions it provokes” (Richards [1929] 1978, 9). North notes that there was a move from the external to the internal when the American “New Critics” embraced Richards’s concepts, and thus the aesthetic became aligned with “the Kantian and idealist realm of transcendental value” (Richards [1929] 1978, 154). Since then, the notion of “aesthetic value” has been connected with the text itself, as disseminated by the “New Critics,” but it should be remembered that Richards initiated a more materialist, external, instrumental approach to the text. Nevertheless, the most problematic aspect of close reading, as Franco Moretti polemically proposes, “is that it necessarily depends on an extremely small canon” (Moretti 2013, 57).

Distant reading is first mentioned in Moretti’s “Conjectures on World Literature” from 2000 in the following way: “Distant reading [...] is a condition of knowledge: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes—or genres and systems” (Moretti 2013, 57). Moreover, it is introduced as a “little pact with the devil” (Moretti 2013, 48): a model for a new kind of approach to comparative literary history that would, in practice, “become ‘second hand’: a patchwork of other people’s research, *without a single direct reading*” (Moretti 2013, 48, original emphasis). This macroscopic perspective draws attention to the fact that world literature is both a single and an unequal system, where the developments at the core are adopted and adapted at the periphery; here, the ideas of the “core” and “periphery” are directly taken from Wallerstein’s economic world systems theory, and their power dynamics applied to the literary sphere. However, the limitations of the proposed approach underline the inequalities of the very system that is being studied, since one could distantly read about literatures only up to the level of formal analysis, when “the study of world literature must yield to the specialist of the national literature, in a sort of cosmic and inevitable division of labour” (Moretti 2013, 59), implicitly assuming that the specialist must be from the periphery. This briefly raises the issue of linguistic competence, but only in passing, ignoring the many complex issues that reading distantly across many different languages and literary traditions entails in practice and completely overlooking the role that translation plays in the transmission of literary trends and cultural ideas. (While we consider that the relationship between multilingualism, multilingual corpora, and literature in translation merits more attention than it has been given so far within the debates on distant reading, it goes beyond the scope of this article. However, it is worth mentioning the efforts of the pan-European COST Action entitled Distant Reading for European Literary History to fill this gap by creating a multilingual European Literary Text Collection [ELTeC], as well as by adapting and developing computational practices to different European literary traditions and by reflecting on the consequences of such approaches. For more on this multilingual endeavour, see https://www.distant-reading.net/.)

This meaning of distant reading shifts in Moretti's article "The Slaughterhouse of Literature" (Moretti 2013), where it moves from reading across national canons towards the even more ambitious goal of looking at "*all* of literary history: canonical and noncanonical: together" (Moretti 2013, 66, original emphasis), even though he never engages in such a multilingual endeavour. The focus now is on an examination of the evolution of genres and canons: which elements, for example, of a detective story, survive to help shape the genre *and* contribute to the recognition of an author's work as canonical, and which ones lead to omission from traditional literary histories? The change of scale now also requires a different set of skills from the ones traditionally employed by historians of literature, such as "sampling; statistics; work with series, titles, concordances, incipits" (Moretti 2013, 67). While this and the subsequent study *Graphs, Maps, Trees* (Moretti 2005) both zoom in on the evolution of genres, neither relies on an extensive use of large-scale digitized corpora. This will only become the steppingstone of the next phase of Moretti's work from 2011 onwards, after the setting up of the Stanford Literary Lab and the publication of the first in a series of its collaboratively produced pamphlets. It was the accelerating digital turn in the study of literature that made the lab's work possible at that particular moment in time, since now the ideas about macroscopic analyses could be tested by using the digitized corpora and the digital tools that were being developed. With the Stanford Literary Lab pamphlets, which have been trying to unsettle the established ideas about the nature of literary archives and canons (Algee-Hewitt et al. 2016) or the very notion of literary interpretation (Moretti 2017), distant reading completed its transformation from a literary history concept that thrived on the formalist-driven tension between evolution and world systems theories, towards a quantitative, large-scale, computer-assisted approach to the study of literature.

As Ted Underwood stresses in "A Genealogy of Distant Reading," such a quantitative turn enables distant reading to "aspire to the methods of social science: it is defined not only by a commitment to historical breadth, but by a version of the scientific method appropriate for a historical discipline" (Underwood 2017, para. 22). Underwood here alerts his readers to both the potential and the risk behind the phrase "distant reading." One of its risks is to erase the work done before the term's introduction by Moretti. Underwood convincingly argues that to coin the phrase should not be equal to inaugurating a new critical method. The critical approach, then, existed well before Moretti first mentioned the term in his "Conjectures." As Underwood posits, "[d]istant reading has evolved into a name for a more specific approach to literary history, but the approach described significantly predates this particular name for it" (Underwood 2017, para. 10). Earlier efforts to manage and process information before the digital age can be dated back to Renaissance Europe, as Ann M. Blair describes in her *Too Much to Know* (Blair 2010). As far as macroscopic literary history is concerned, Underwood pushes the departure point for Distant Reading back to the mid-twentieth century with critics such as Raymond Williams and Janice Radway who "worked on the boundary between literary history and social science" (Underwood 2017, para. 5). For him, distant reading should not be conflated with data research or computer-based analysis, as it has to do with "[i]ntegrating experimental inquiry in the humanities" (Underwood 2017, para. 5). And, in his view, this was achieved in the 1980s and in the 1990s by critics like Radway in her 1984 book *Reading the Romance*. In the introduction to this book, she expounded her method and argued that she saw "necessary to connect particular texts with the communities that produced and consumed them and to make some effort to specify how the individuals involved actually constructed those texts as meaningful semantic structures ... what American Studies needed were ethnographies of reading" (Radway [1984] 1987, 4). Therefore, Radway was already opening up a path later to be trodden by Moretti and his followers in understanding her endeavour as an experimental enquiry.

Seen in this light, distant reading is concerned with new methods of large-scale literary history, and in so doing, it "centres on a social-scientific approach to the literary past" (Underwood 2017, para. 41). However, as Katherine Bode has astutely pointed out, what has often been overlooked in the discussions about distant reading, as well as macroanalysis, is the way that the meaning of both terms changed over time, so that even though both "data and computation remain central, the primary object of distant reading is now less often literary-historical systems—particular social, material, and political contexts for literary development and change—than 'the concepts of literary study'" such as "characterization, plot, and dramatic form" (Bode 2017, 79–80). It is this understanding of distant reading that will be in focus here: as a set of approaches to the study of literature enabled by computer-assisted analyses of large-scale corpora, made possible by the digital turn in the humanities and aimed at the examination of literary concepts. The following sections will therefore examine the ways in which it has impacted the conceptualization of genre, authorship, and style, before reflecting on its place within the broader context of the digital turn in the humanities.

2. In focus: Genre, authorship, style

2.1. The digital turn and the concept of genre

2.1.1. Naming literary genres

Before the pragmatic shift caused by the distant reading approaches, a similar shift in the theory of (literary) genres occurred under the influence of the new rhetoric, chiefly of *situationism*: "people use genres to do things in the world ... genres are defined less by their formal conventions than by their purposes, participants, and subjects" (Devitt 2000, 698); thus "genres are never purely literary" (Compagnon 2001, sec. 3). Following a path opened by the distinction "intrinsic"/inner *vs* "extrinsic" conditional genres (Hirsch 1967; Genette et al. 1986), scholars tried to leave behind the normative, taxonomic, and perhaps static meaning of this concept and adopt a descriptive, empirical, and situation-driven viewpoint (Wellek and Warren [1942] 1967). New readings of Aristotle—Hamburger's "logic of enunciation" (Hamburger [1957] 1986) and Genette's theory of arch-text are the most representative—delved into the self-perpetuating inflation of genres and emphasized that instead of "genres" we should use "modes" (Genette [1979] 1994, 67–80). Indeed, the dyads "fiction"–"diction" and "fictional/ mimetic"–"lyric" attempted to renovate what used to be esteemed as *natural* and *given-by-default* in the traditional system of genres, by erasing the generic limits, "the law of genre" as Derrida calls it (Derrida 1980, 56–57), by expanding each genre's transformative possibilities and, as if a new label could actually change the content, by renaming something regarded as obsolete. Hamburger and Genette argue that genres do not belong within the world of literary symptoms—and genre inflation should be treated as a symptom, but with an immanent logic of creation whose *primum movens* lies in the language itself rather than in the natural order or in the human psyche. However, if we credit Jean-Marie Schaeffer, Aristotle and his modern followers champion a kind of hardcore essentialism which boils down to the fact that genres are filled with a substance; that generic labels are ontologically anchored signs and not just conventions (Schaeffer [1989] 2006).

Like Schaeffer, Tomașevski discovered the paradox in naming genres: generic labels can be extremely conservative, and they might be erroneously correlated with contents that have changed in the meantime (Tomașevski [1925] 1973). It is in this line of anti-essentialist thought that we should thus integrate Tomașevski's own analogy *genre-genetics* (Tomașevski [1925] 1973), Mikhail Bahtin's intuition on *finite* and *infinite* genres (Bahtin [1975] 1982), Tzvetan Todorov's and Fredric Jameson's ecumenical ideas that genres must be "the meeting place between the general poetics and event-based literary history" (Todorov [c. 1978] 1990, 19–20) or "a coordination of immanent formal analysis of the individual text with the twin diachronic perspective of the history of forms and the evolution of social life" (Jameson 1981, 92), and many other recent works that endeavour to address what John Frow calls "the madness of genre" (Frow 2007, 1627).

2.1.2. Generic wateriness: Medium, schema, database, possible world

"The mediatory function of the notion of genre" (Jameson 1981, 92), the possibility to think of *genre* as a median line or even as *medium* between other literary institutions has cured literary theory of "the mystique of literary genres" (Lovinescu [c. 1928] 1981, 398–401) and pushed it outside the circle of "magical thinking"; yet the frequent use of a generic label, Schaeffer notices, does not warrant the existence of a corresponding entity (Schaeffer [1989] 2006). The genre is thus a lively *mediator* between a tolerable level of literarity and highly praised originality, between literature and its masterpieces (Compagnon 2001), a *medium* that, according to McLuhan's catchphrase, is the same with the message, a *mediating structure* between texts/libraries/databases/corpora and situation/circumstance/context. Approached accordingly, "the use-value" of genres, understood as usefulness and usability, is embarked upon its new career in distant reading (Devitt 2000, 707). Thereupon, the name of the genre, whatever the users might choose as fit for them in a given situation, actually means the genre itself: "Genres have solid names, ontologized names. What these names designate ... is not taxonomic classes of equal solidity but fields at once emerging and ephemeral, defined over and over again by new entries that are still being produced" (Dimock 2007, 1379).

Indeed, with the massive advent of digitization and with the development of computational literary studies, the new input from rhetoric, cognitive sciences, anthropology, microsociology, sociology of classification, computer studies, and, last but not least, possible-world theory, has turned the concept into one of the most fascinating areas of distant reading research. "We could think of genres," John Frow believes, "as clusters of metadata ... that help define the possible uses of textual materials" (Frow 2007, 1631). By grafting the cognitivist *schema* and *possible worlds theory* on the old notion of "genre," Frow stresses the heuristic value of this new device:

> Genre cues act rather like context-sensitive drop-down menus in a software program, directing me to the layers and sublayers of information that respond to my

> particular and local purposes as speaker, reader, and viewer … [F]ar from being merely stylistic devices, genres create effects of reality and truth that are central to the ways the world is understood. (Frow 2007, 1631–1632)

In a framework organized by the multiple use of "the genre database" on the one hand, and by the theory of *world literature* on the other, Wai Chee Dimock defines genres as "fields of knowledge," as "open sets endlessly dissolved by their openness … virtual in this nontechnical sense, resembling the database in being an unscripted effect of their membership and in being only a fraction of what they could be at any given moment" (Dimock 2007, 1379). Hence, we could read Dimock's take as a reinterpretation of Todorov's and Jameson's integrative view, as "a combination of a long-time frame and a short observational distance" (Dimock 2007, 1382) where this new take on "the generic wateriness" (Dimock 2007, 1379) rather than on self-contained genres could introduce a new type of literary pedagogy. But apart from pedagogy—that is, changing the way literature is taught and studied—is there any other reason for genres to be read distantly?

Seen from afar, and therefore as virtual and not as actual, genres are "stackable," "switchable," and "scalable" (Dimock 2007, 1379). This user-oriented theoretical frame does not break up with the old matrix of "expectation and recognition" (*modèle d'attente et de reconnaissance*) that has always been the groundwork of genres (Compagnon 2001, sec. 2) but rebuffs the romantic infatuation with originality, masterpieces, and exemplary figures. At the same time, it is obvious that stackability, switchability, scalability (and other features that may be drawn from the image of "wateriness") refer more to what is corresponding (in terms of binary-computed data) to an arbitrary or subjectively given generic label, and less to a fixed, thus recognizable, name. A genre's name—be that name "tragedy," "epic poem," "comédie larmoyante," "city mystery," "drawing room play," "melodrama," "idyll," "hajduk fiction," "French novel," "Spanish novel," "silver-fork novel," or else—is already an assumption; thus, it comes with a heritage of commonplaces, presuppositions, and critical practices. Still, generic labels are something to invent and use with caution.

2.1.3. Back to genre: Matrix, topic model, subgenre, microgenre, folksonomy

In 2011, when the first Stanford Literary Lab Pamphlet was published, the names of genres created a bit of vexation among the team members: "right now, the very names of novelistic genres are a telling—even maddening sign—of categorical confusion highlighting now the novel's medium (the epistolary novel), now its content (historical, industrial), style (naturalist), protagonist (picaresque, pastoral), all the way to more or less fanciful metaphors (Gothic, silver-fork)" (Allison et al. 2011, 10). Yet, they have not lost either hope or scope of research, which, back in those days, read as follows: "the system of genres might turn from a hodge-podge of unrelated categories to a single matrix of interconnected formal values" (Allison et al. 2011, 10). Gothic, historical, industrial, silver-fork, anti-Jacobin, evangelical, Newgate, Jacobin, and sensation novels plus national tales and Bildungsroman, this was the menu on the researchers' plate. Because "language and style are not just enough to delimit one genre from another," and because nineteenth-century novels (the Stanford corpus) rely rather on plot than style, individual literary styles looked like a more reachable target than genre. Statistical findings proved that genres were "icebergs: with a visible portion floating above the water, and a much larger part hidden below, and extending to unknown depths" (Allison et al. 2011, 25), which must have felt reassuring for traditional genre theorists. However, the pamphlet's authors suggested that their bootstraps and consensus trees also indicated "a conflict of forms." Stanford researchers discovered that, like the authors launched on the literary market (Moretti 2013), "genres engage[d] in a struggle for recognition" (Allison et al. 2011, 18) and acted as rivals.

While the "iceberg" metaphor could encourage, in Tomaşevski's line, some theoretical musings on dominant/seen and recessive/unseen features, the picture of genres fighting for supremacy is somehow disquieting, as, in accepting this underlying personification, we actually accept and bounce back to the essentialist-evolutionist perspective. It is neither good nor wrong to see "genre" more as a *prescription*, more as an *essence*, or more as a *structure* (Compagnon 2001, sec. 3). It is strange, however, to argue—as Moretti does—that

> usually, we tend to have a rather Platonic idea of genre: an archetype and its many copies (the historical novel *Waverley* rewritten over and over again; the picaresque Lazarillo and his siblings). The tree suggests a different image: branches, formal choices, that don't reduplicate each other but rather move away from each other, turning genre into a wide field of diverging moves and wrong moves mostly. (Moretti 2013, 71)

In the vein of his remarks, it is probably true that usually "we" tend to have a rather Platonic approach to names of genres too. The truth is that Moretti's *Distant Reading* circulates a very traditionalist and top-down approach to genre. "The text" (whatever *text* might mean in Moretti's mind) stands between two formal units: "the device" and “the genre” (Moretti 2013, 77). A mention of “subgenre” indicates that, in this particular analysis, genre labels (and their load of commonplaces, presuppositions, and critical practices) have been stronger than computationally modelled data: “large genres like tragedy, or the fairy tale, or even the novel, seem rooted in the *longue durée*, while ‘subgenres’ (the gothic, the silver-fork school, the Bildungsroman, the nautical tale, the industrial novel, etc.) thrive for shorter periods (thirty to fifty years, empirical findings suggest)” (Moretti 2013, 86).

A different relationship between the generic label and its corresponding content can be found in Ted Underwood’s publications from 2013 on. In dealing with big data (Hathi Trust digital library), with the critics’ list of generic labels, and with the prevailing nominalism in the theory of genre, Underwood’s approach to genre was a bottom-up and agnostic one:

> Many research teams are creating collections manually, selecting novels or works of poetry one by one, guided by existing bibliographies.... Instead of building a new collection each time we add a genre, period, or problem to our research agenda, we could be defining and redefining collections simply by selecting subsets of the library. (Underwood 2014, 4)

An algorithmic approach to genre meant to stabilize its “ontology,” which focused research on “categories that have visible formal characteristics (a lyric poem or an index)” (Underwood 2014, 7): drama, fiction, non-fiction, poetry, and paratext; more precisely, the typical “page” of each of the five large “genres.” The results have shown that “narrower generic categories tend to be less stable,” and that “as we move into narrower subgenres within these divisions, genre may more closely resemble a folksonomy” (Underwood 2014, 9). Consequently, what a wise theorist of genres can do is to doubt everything and rely on predictive rather than explanatory models:

> An explanatory model attempts to identify the key factors that cause or define a phenomenon. A predictive model doesn’t claim to reproduce this sort of deep structure.... [I]t starts by accepting the phenomenon to be modelled as something whose internal logic is “complex and unknown.” Instead of attempting to capture the original causal logic of the phenomenon, it looks for an adequate substitute: a function that maps predictor variables onto response variables in a roughly equivalent way. (Underwood 2014, 10)

The predictive attitude concerning data represents the basic principle of predictive modelling, from David Blei’s analysis of *Science* articles (Blei 2012) to Jonathan

Goodwin's recent approach to "machine-classified microgenres" (Goodwin 2020), of authorship attribution and stylometry. Indeed, why speculate when prototyping has a greater force of argumentation (Galey and Ruecker 2010)? In fact, scholars have recently emphasized that all genres, as well as all clusters that are presumed to send out genre signals, must be re-modelled or at least rechecked from the perspective of conventional genre distinctions and typical topic patterns (Schöch 2017).

In the broader framework of distant reading, the shift from explanatory to predictive approaches to genre might turn the concept into a user-oriented tool able to track down the "visible" parts of texts and devise new ways of clustering them.

2.2. Authorship and the digital turn

In the twentieth century, the concept of the author underwent a radical redefinition, playing a pivotal role in the studies of language, writing, and meaning. Overall authorship studies aim to detect common features in a single or a corpus of texts and create a pattern of authorship markers. Authorship attribution has a long history of using a large variety of textual features in order to correlate them with a specific author's style. During the last couple of decades, this scientific field has been developed substantially by taking advantage of research advances in areas such as machine learning, information retrieval, and natural language processing (cf. Stamatatos 2009). Some of these developments in authorship attribution have also drawn the attention of popular media as well as that of the wider academic community, most prominently through their interventions in the debates on the authorship attribution of Shakespeare's works (see, e.g., Vickers 2011). Indeed, the identification of the authors of the Gospels as well as the authorship attribution in the case of Shakespeare, Marlowe, and others; collaborative authorship; the scope and degree of an author's authority; the role of authorial intention and biographical and autobiographical information in interpretation—these are all issues that have been discussed with a vigour that testifies to the high stakes of the authorship question (Donovan, Zadworna-Fjellestad, and Lundén 2008).

The authorship attribution approach has a long history and a wide range of applications, and determining the author of a particular piece of a text has raised methodological questions for centuries. According to Holmes (Holmes 1994), authorship attribution dates at least to more than a century ago with the work that proposed distinguishing authors by looking at word lengths (see, for example, Mendenhall, 1887). This was later improved by Yule (1939), where the average length of sentences was considered as a determinant. A seminal development was the introduction of the analysis of function words to characterize authors' styles on the authorship of the disputed Federalist Papers (Mosteller and Wallace 1964), which inspired the development of several methods.

Overall, authorship attribution is defined as the science of inferring characteristics of the author from the characteristics of documents written by that author (Juola 2008). As several studies in intellectual property rights have shown, there is a close relationship between literary property laws and the cultural construction of authorship. In fact, as Mark Rose observes, copyright is founded on the notion of original authorship (Rose 1993). Therefore, questions of authorship can be of interest not only to humanities scholars, but in a much more practical sense to politicians, journalists, and lawyers. Investigations and simple close reading by experts have traditionally given good results. However, recent developments in distant reading approaches that combine improved statistical techniques with the wider availability of computer-accessible corpora have made the automatic and objective inference of authorship a practical option.

The issues of how the term "author" should be defined and what the functions of an author are have been a major concern for over half a century for philosophers, theorists, and writers, from Coleridge and Eliot through Benjamin, Barthes, and Foucault to Derrida. Barthes proclaimed the "Death of the Author" in 1967 and spoke about the idea of "authorship" as just a cultural convention that reflects the capitalist ideas of ownership and individual prestige. In 1969, Foucault took the author debate further, studying how the notion of "author" is related to the author's function. He highlighted the fact that every text is made of multiple texts that have preceded it. Dusollier underlines that thanks to the reconsideration of the author in the works of Foucault and Barthes, the notions of "hypertext" and "intertextuality" emerge from the new vision of a text as having an "evolutionary, modifiable and open nature" (Dusollier 2003, 290) and the notion of the public as being the actual author of the work. Dusollier considers these notions as "the founding principles of both free software and free art movements" (Dusollier 2003, 282), which she relates to Web 2.0 and to the new collaborative culture shared by internet users worldwide. Foucault and Barthes brought the role of intertextual forms of meaning creation to the forefront; thus, quantitative authorship attribution in literary texts has had to deal with the epistemological status of authorship itself (Herrmann, Van Dalen-Oskam, and Schöch 2015).

When it comes to defining the features of authorship attribution, as Sari, Stevenson, and Vlachos point out, these have been divided by Stamatatos into five groups, namely:

> lexical, character, syntactic, semantic and application-specific features. Compared to others, lexical and character features are commonly used in authorship attribution work as they provide rich information about the author's topical preferences

> and writing style. In addition, both types of features can be extracted in many languages and datasets with little effort. (Sari, Stevenson, and Vlachos 2018, 344)

However, Juola has observed at least three main problems in authorship attribution:

> The first is, given a particular sample of text known to be by one of a set of authors, determine which one. This, “closed class,” version of the problem is closely related to the second, “open class,” version of the problem: given a particular sample of text believed to be by one of a set of authors, determine which one, if any, or even here is a document, tell me who wrote it. The open-class version is of course much harder to solve, especially if one is required to distinguish among people outside of a small candidate set. The third problem—for which some researchers prefer to reserve the word “stylometry” or “profiling,” reserving “authorship attribution” only for the first two—is that of determining any of the properties of the author(s) of a sample of text. (Juola 2008, 238)

In many areas of the humanities that rely on traditional textual media, the distributed author is alive and remains a current object of study according to the premise that a so-called pre-modern society was less concerned with individuality and notions of property were unknown or at least insignificant. For example, the notion of “distributed authorship” supports Classical antiquity studies, with Homer as its foundational point of orientation. In recent years, the dynamic possibilities of distributed authorship have accelerated most rapidly in media associated with the virtual domain, where modes of communication have rendered artistic creation increasingly collaborative, multi-local, and open-ended. As Simone Murray has pointed out,

> Sectors of the contemporary digital literary sphere have attempted to offset or counteract wholesale democratization of authorship via a variety of means: from the avant-gardist self-stylings of experimental electronic literature authors; through the editorial overseers of collaboratively written wiki or remix projects; to the perversely hierarchical beta-reader protocols prevalent in fanfic communities. (Murray 2019, 52)

In addition, during the twentieth century the authorship attribution came to be closely associated with the more general and varied practice of “stylometry,” and it remains an important aspect of text analytics today (Schreibman, Siemens, and Unsworth 2015); the application of distant reading approaches, based on the application of statistical analysis to large-scale corpora, has also contributed to the development of new quantitative methods. Authorship attribution is a unique task closely related to the representation of individuals' writing style and text categorization; thus, the changing definition of the concept of style will be the topic of the next section.

2.3. The digital turn and the concept of style

As it has been pointed out, authorship attribution with computational means has long been synonymous with "stylometry," also called "computational stylistics." A brief history of stylistics before stylometry in Europe necessarily revolves around the evolution of the concept of style, understood as the manner of writing or speaking (*modus scribendi/dicendi*, cf. Sowinski 2013). All the way from antiquity, it was a central category of rhetoric, described in the teachings of *elocutio*, the mastery of stylistic elements, which had the ultimate goal of audience persuasion. Here, a particular level of style was to be matched to an intended effect. Scholars studied high, low, and middle style by authorial examples. The notion was largely prescriptive for long, but took a descriptive turn around 1900 when "positivist" and "formal" studies of literary prose started to flourish. A strong tradition in style research sees a relationship between language use and an author's psychology, with eminent proponents such as Buffon, Dilthey, the early Leo Spitzer, and René Wellek and Austin Warren (Buffon [1753] 2007; Dilthey 1887, esp. chap. II.2; Spitzer [1928] 1931; Wellek and Warren [1942] 1956, esp. chap. 8). They have most recently been followed up within an empirical paradigm by psychological computational text analysis, mapping author's word use onto different psychological constructs, such as thought, emotions, and behaviour (Boyd 2017).

Raising a need for the externalization of scholarly assumptions and formalization of method, the digital turn prompted scholars to revisit the definition of style. Herrmann, Van Dalen-Oskam, and Schöch examined the approaches to "style" in French, German, and Dutch literary studies since 1945 and observed that a formal methodology, especially one using measures of word frequencies, has hardly played a role (Herrmann, Van Dalen-Oskam, and Schöch 2015). By contrast, stylometry (today largely synonymous with computational stylistics) potentially considers all words in the texts under consideration and does so from a quantitative vantage point. Stylometry has long mainly focused on authorship, discussed above; one of its key findings is that the true (i.e., most probable) author of a text is most visible in the frequency patterns of the most frequent words, which are typically function words, for example "and," "the," "she," "our," "in," etc.

Before the digital turn, function words generally were ignored—with the exception of John Burrows's dissertation *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*, which can be seen as the start of and inspiration for much of today’s work in stylometry and computational literary studies (Burrows 1987). Over time, stylometry has had a marked tendency to increase the number of words that are considered: from several dozen most frequent function words in pioneering work of the 1960s to the many hundreds or even several thousand different most frequent words, including many thematic terms, employed in recent stylometric work. From a consideration of a limited number of specific words or phrases, then, the object of stylistics has changed to function words and large parts of the vocabulary of a text or corpus, driven by methodological and technical advances. To include this new, distant reading view on literary texts, Herrmann, Van Dalen-Oskam, and Schöch came up with a new definition of style: “Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively” (Herrmann, Van Dalen-Oskam, and Schöch 2015, 44). This definition aims at establishing a minimal common ground for descriptive style studies, deliberately allowing for both distant and close reading. As far as we can tell, this definition of style has been received well by the distant reading community, if the ways in which it has been adopted by fellow researchers is any indication. (One place allowing scholars to retrace the reception of this definition is, for want of a better infrastructure, the Google Scholar entry for the original article: https://scholar.google.com/scholar?hl=en&cites=11491311335836562153.)

Quantitative approaches to style today often go beyond authorship and search for ways to compare the style of different works, authors, genres or subgenres, periods, etc. by applying statistical methods to all words used in the texts, or to analyze such features as sentence length, vocabulary complexity (lexical density), and sentence complexity. Topic modelling may be used to find thematic patterns. Sentiment analysis, quantifying word use that may be linked to positive and negative emotions, is used to search for narrative patterns throughout texts (cf. Herrmann, Jacobs, and Piper 2021). Quantitative approaches to style are also used in combination with results from reader surveys, in search of which textual and contextual elements may play a role in readers’ perceptions of literary works (Koolen et al. 2020; Van Dalen-Oskam 2023).

Distant reading today is able to investigate larger and larger collections of literary texts that are becoming available, and to use appropriate software for text processing, machine learning, visualization, and statistics (e.g., Python, R, Java with dedicated packages or libraries such as scikit-learn, stylo, or MALLET). Using these datasets and tools, distant reading of style today focuses on the occurrence of textual patterns defined at various levels, from characters and single tokens to n-grams and larger, more complex, and more abstract linguistic structures or features. It often focuses on the various ways that these patterns interact with contextual factors, such as authorship, author gender, period of publication, or literary genre at the document level (and as available to the analysis via document-level metadata) as well as dialogue and narrative or stage direction vs. character speech at the textual level (available to the analysis via inner-textual annotations or pattern finding methodology). It also frequently focuses on the relationship of such patterns with a range of potential functions of literature that can be aesthetic, communicative, social, as well as cognitive. As these multiple avenues of research indicate, analysis of style in the distant reading paradigm has become a wide-ranging, multi-faceted research activity.

To sum up: new methodological developments led to a new practice of stylistic analysis, in turn prompting a new definition of style. The new approaches to style inspired a change of focus in research. Instead of focusing on only a small selection of salient stylistic features which can be analysed "manually" and occur in relatively low frequencies, the new computational methods have enabled the analysis of many or even all words in the texts. The most frequent words occur so often that their usage, and the variation of their usage in different texts or groups of texts, could not have been grasped by earlier researchers. Now that this can be done, researchers have started to explore as of yet uncharted parts of (literary) language. This includes types of style analyses that extend to trans-textual perspectives, allowing questions about genres, periods, and other larger units of scholarly interest. The distant reading of style reveals patterns in language usage (for example, of function words including pronouns, articles, prepositions, adverbs, and others), which may lead to a much more detailed knowledge of the art of writing, reading, and reasoning.

3. Conclusion: Distant reading and the digital turn

As the three discussions of literary concepts above have shown, the effect of the distant reading approaches on the study of literature has been substantial, challenging the received notions of genre, authorship, and style, and contributing to their reconceptualization. By way of conclusion, it is also important to point out that the development of distant reading as a large-scale, computer-assisted, approach to the study of literature has not been an isolated phenomenon. Since the mid-1990s a series of concepts and digital approaches have disrupted the humanities. The opening up of archives and collections through digitization, for example, has allowed historians to complement existing methodologies with a more macroscopic understanding of the past. Digital technologies allowed historians to interrogate their sources more easily at a range of interlocking scales (see Guldi and Armitage 2014; François et al. 2016). Such a move towards grappling with research questions at a range of scales and with the use of multiple, largely digital, sources and methodologies can also be seen in the history of art—where the term "distant viewing" has also been gaining traction of late. Whereas it is important to acknowledge the uneven impact of the digital turn across the humanities and even within specific fields, it is nevertheless clear that this engagement with digital methodologies has been immensely fruitful. It is therefore useful to understand what these digital turns for each of the humanities disciplines have in common with one another and how the impact of distant reading can be better understood through this lens. For this purpose, we here offer some reflections on some important commonalities between distant reading and other areas of the digital humanities, and the shared challenges brought about by the digital turn, before turning to the challenges particular to distant reading.

The trajectory in terms of what the digital promises and what it actually delivers seems to follow a similar pathway across different humanities disciplines. The introduction of disruptive concepts, underpinned by digital technology, holds out at first the promise of radical change. This promise is usually underpinned by a phase during which various overlapping definitions are being proposed and during which a lot of intellectual effort is invested in theoretical reflections. Some empirical work is also undertaken but this is not always fully integrated with the wider theoretical debates. Only in a second phase the empirical work takes the main stage, and debates about definitions start to play out at a less intense level. The complexity of the results and the difficulty to interpret these leads to an increased reconnection with the traditional humanities, which the first phase had often weakened. From the perspective of 2022, many of the new methodological challenges are now viewed as being more closely aligned with older existing methodological challenges. The trajectory of the development of distant reading from a theoretical concept to a series of methodological approaches, as detailed in the introduction, broadly fits this observation. However, it is equally important to ensure that there is a good flow of lessons learned and models of good practice between the different areas of the digital humanities. Although the domain-expertise and the choice of methods often differs greatly amongst digital humanists, the new digital methodologies do share similar challenges.

The most common shared challenge is posed by the simplistic equation of the "digital" with a clear-cut data science approach and the ability to ask new, and often bigger, questions, since this is an unhelpful reduction of a much more complex and vibrant reality. Many digital humanists, across all disciplines, are at pains to highlight that their questions have been lying at the core of the humanities disciplines and that they had important and influential advocates in the field long before the digital turn. In the context of distant reading, this point has already been made forcefully earlier in this article; when it comes to historical research, for example, the desire to quantify the human past, although greatly facilitated by the digital turn, has been part of the historical discipline for decades. Furthermore, the digital turn should not be narrowed down so that it means the same thing as digitization. The digitization of collections is an absolutely vital part of the digital ecozone but the "digital" impacts also the selection of material, the analysis, and the presentation of the results, and while it enables the application of computer-assisted methods, these do not exclude or make redundant non-computational approaches.

Zooming back in on the effect of the digital turn on the study of literature, we can see similar effects at work. The change of scale that digitization of the literary archive brought about, alongside the development of digital tools for textual analysis, has enabled the transformation of the rather abstract and fairly provocative concept of "distant reading" into a series of computer-assisted quantitative approaches. In turn, the application of these to the study of genre, authorship, and style changed not only the way these have been studied and analyzed, but also the way they have been defined, putting both old as well as new definitions of these concepts to the test. The current main challenge is to integrate carefully the results of large-scale, computer-assisted distant reading studies within older existing bodies of literature, in order to establish which frameworks, concepts, and methods stand, need modification, or can be rejected.

While the digital turn has made it possible to, quite literally, think big in terms of the size and number of the corpora that could now be considered for literary analysis, several other challenges remain. The first and the most basic one is that of unequal access to digitization across the world, underlining the point made earlier that digital turn is not the same as digitization. If digitization of national literary archives is the sine qua non of distant reading and a prerequisite for the development of computer-assisted study of all world literatures, either individually or comparatively, then unequal access to digitization and digital tools works to reproduce the kind of world-systems inequality between the (Anglophone) "core" and (non-Anglophone) "peripheries," that Moretti's early conceptualization of distant reading (Moretti 2013), discussed earlier, was, at least nominally, meant to redress. While in some countries their national literary archives have been extensively digitized, in others the process has barely begun; as a consequence, any attempt at comparative literary study, such as the mapping of the geographical spread of literary genres, for example, needs to treat the lack of data as such, rather than using it to make assumptions about the literary traditions in question (Primorac 2023).

The other major remaining challenge is also the result of the oversight mentioned at the start of the article. When applying distant reading as a computer-assisted *method* to test distant reading's *theoretical* assumptions about world literature as a single yet unequal literary system, one encounters the practical problem of language. Namely, it is not just that the current tools for textual analysis still do not allow for the kind of multilingual textual analysis that would be necessary to produce computer-assisted world literature studies that could compare more than two to four literary traditions at one time—the number, accidentally, not much bigger than what is considered standard in traditional comparative literature analysis, criticized by Moretti as limited, if not limiting, in traditional comparative literature studies (Moretti 2013). The fact of the matter remains that tools still need yet to be developed for a number of small languages; as a consequence, authorship attribution analyses or stylometric analyses of literary texts can only be done on and theorized about within a limited number of literary traditions. If distant reading, macroanalysis, and data-rich literary history in general were to truly live up to their promise of opening up the world literary archive to new modes of computer-assisted literary analysis, comparative or not, then both the digitization and the digital tools also need to become available beyond the (mostly Anglophone) core. Otherwise, the distant reading approach to the study of world literature will remain forever locked into the unequal, if single, system of power relations between the culturally powerful "core" and the globally dispersed "peripheries."

Competing interests

The authors have no competing interests to declare.

Contributions

Authorial

Authorship is alphabetical after the drafting author and principal technical lead. Author contributions, described using the CASRAI CredIT typology, are as follows:

Author name and initials:

Rosario Arias (RA)
Eva Eglāja-Kristsone (EEK)
Pieter François (PF)
Berenike Herrmann (BH)
Roxana Patras (RP)
Antonija Primorac (AP)
Christof Schöch (CS)
Karina van Dalen-Oskam (KDO)

Authors are listed in descending order by significance of contribution. The corresponding author is AP:

Conceptualization: AP, RA, PF, KDO, RP, BH, CS, EEK
Writing – Original Draft Preparation: RA, EEK, PF, BH, RP, AP, CS, KDO
Writing – Review & Editing: AP

Section contributions are as follows:

1. Introduction (AP and RA)
2.1. The digital turn and the concept of genre (RP)
2.2. Authorship and the digital turn (EEK)
2.3. The digital turn and the concept of style (BH, KDO, CS)
3. Conclusion (PF and AP)

Editorial

Section and Layout Editor

A K M Iftekhar Khalid, The Journal Incubator, University of Lethbridge, Canada

Copy Editor

Christa Avram, The Journal Incubator, University of Lethbridge, Canada

References

Algee-Hewitt, Mark, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and **Hannah Walser.** 2016. “Canon/Archive: Large-Scale Dynamics in the Literary Field.” *Stanford Literary Lab Pamphlet* 11. January. Accessed November 10, 2022. https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf.

Allison, Sarah, Ryan Heuser, Matthew Jockers, Franco Moretti, and **Michael Witmore.** 2011. “Quantitative Formalism: An Experiment.” *Stanford Literary Lab Pamphlet* 1. January 15. Accessed November 10, 2022. http://litlab.stanford.edu/LiteraryLabPamphlet1.pdf.

Bahtin, Mihail. (1975) 1982. *Probleme de literatură și estetică*. Translated by Nicolae Iliescu. Bucharest: Univers.

Blair, Ann M. 2010. *Too Much to Know: Managing Scholarly Information before the Modern Age*. New Haven: Yale University Press.

Blei, David. 2012. “Probabilistic Topic Models.” *Communications of the ACM*, 55(4): 77–84. Accessed November 10, 2022. http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf. DOI: https://doi.org/10.1145/2133806.2133826.

Bode, Katherine. 2017. “The Equivalence of ‘Close’ and ‘Distant’ Reading; or, Toward a New Object for Data-Rich Literary History.” *Modern Language Quarterly* 78(1): 76–106. DOI: https://doi.org/10.1215/00267929-3699787.

Boyd, Ryan L. 2017. “Psychological Text Analysis in the Digital Humanities.” In *Data Analytics in Digital Humanities*, edited by Shalin Hai-Jew, 161–189. Cham: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-54499-1_7.

Buffon, Georges Louis Leclerc. (1753) 2007. “Discours sur le style.” In *Oeuvres*, edited by Stéphane Schmitt. Bibliothèque de la Pléiade. Paris: Gallimard.

Burrows, John Frederick. 1987. *Computation into Criticism: A Study of Jane Austen’s Novels and an Experiment in Method*. Oxford: Clarendon Press.

Compagnon, Antoine. 2001. *Théorie de la littérature: la notion de genre*. Accessed November 10, 2022. https://www.fabula.org/compagnon/genre.php.

Derrida, Jacques. 1980. “The Law of Genre.” *Critical Inquiry* 7(1): 55–81. https://www.jstor.org/stable/1343176.

Devitt, Amy J. 2000. “Integrating Rhetorical and Literary Theories of Genre.” *College English* 62(6): 696–718. DOI: https://doi.org/10.2307/379009.

Dilthey, Wilhelm. 1887. “Die Einbildungskraft des Dichters: Bausteine für eine Poetik.” In *Philosphische Aufsätze*, Vol. 10, 303–482. Leipzig: Fues. Accessed June 26, 2023. https://www.deutschestextarchiv.de/book/show/dilthey_poetik_1887.

Dimock, Wai Chee. 2007. “Introduction: Genres as Fields of Knowledge.” *PMLA* 122(5): 1377–1388. https://www.jstor.org/stable/25501790. DOI: https://doi.org/10.1632/pmla.2007.122.5.1377.

Donovan, Stephen, Danuta Zadworna-Fjellestad, and **Rolf Lundén,** eds. 2008. *Authority Matters: Rethinking the Theory and Practice of Authorship*. Vol. 43 of *DQR Studies in Literature Online*, series edited by Eva Zettelmann and Sylvia Mieszkowski. Amsterdam: Rodopi.

Dusollier, Severine. 2003. "Open Source and Copyleft: Authorship Reconsidered?" *The Columbia Journal of Law & the Arts* 26: 281–296. Accessed April 11, 2023. https://ssrn.com/abstract=2186190.

François, Pieter, Joseph Gilbert Manning, Harvey Whitehouse, Rob Brennan, Thomas Currie, Kevin Feeney, and **Peter Turchin.** 2016. "A Macroscope for Global History: Seshat Global History Databank, a Methodological Overview." *Digital Humanities Quarterly* 10(4): 13. Accessed November 10, 2022. http://www.digitalhumanities.org/dhq/vol/10/4/000272/000272.html.

Freedman, Jonathan. 2015. "After Close Reading." *The New Rambler*. Accessed November 10, 2022. http://newramblerreview.com/book-reviews/literary-studies/after-close-reading.

Frow, John. 2007. "'Reproducibles, Rubrics, and Everything You Need': Genre Theory Today." *PMLA* 122(5): 1626–1634. DOI: https://doi.org/10.1632/pmla.2007.122.5.1626.

Galey, Allan, and **Stan Ruecker.** 2010. "How a Prototype Argues." *Literary and Linguistic Computing* 25(4): 405–424. DOI: https://doi.org/10.1093/llc/fqq021.

Genette, Gérard. (1979) 1994. *Introducere în arhitext. Ficțiune și dicțiune.* Translated by Ion Pop. Bucharest: Univers.

Genette, Gérard, Hans Robert Jauss, Jean-Marie Schaeffer, Robert Scholes, Wolf Dieter Stempel, and **Karl Vietor.** 1986. *Théorie des genres*. Paris: Seuil.

Goodwin, Jonathan. 2020. "Machine-Classified Microgenres." In *The Microgenre: A Quick Look at Small Culture*, edited by Molly C. O'Donnell and Anne H. Stevens, 189–195. London: Bloomsbury Academic. DOI: https://doi.org/10.5040/9781501345845.ch-021.

Guldi, Jo, and **David Armitage.** 2014. *The History Manifesto*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781139923880.

Hamburger, Käte. (1957) 1986. *Logique des genres littéraires.* Translated by Pierre Cadiot. Paris: Seuil.

Herrmann, J. Berenike, Arthur M. Jacobs, and **Andrew Piper.** 2021. "Computational Stylistics." In *Handbook of Empirical Literary Studies*, edited by Donald Kuiken and Arthur M. Jacobs, 451–486. Berlin: De Gruyter. DOI: https://doi.org/10.1515/9783110645958-018.

Herrmann, J. Berenike, Karina van Dalen-Oskam, and **Christof Schöch.** 2015. "Revisiting Style, a Key Concept in Literary Studies." *Journal of Literary Theory* 9(1): 25–52. DOI: https://doi.org/10.1515/jlt-2015-0003.

Hirsch, Eric Donald. 1967. *Validity in Interpretation*. New Haven: Yale University Press.

Holmes, David I. 1994. "Authorship Attribution." *Computers and the Humanities* 28: 87–106. https://www.jstor.org/stable/30200315. DOI: https://doi.org/10.1007/BF01830689.

Jameson, Fredric. 1981. *The Political Unconscious: Narrative as a Socially Symbolic Act*. London: Routledge.

Jockers, Matthew. 2013. *Macroanalysis: Digital Methods and Literary History*. Champaign: University of Illinois Press. DOI: https://doi.org/10.5406/illinois/9780252037528.001.0001.

Juola, Patrick. 2008. "Authorship Attribution." *Foundations and Trends® in Information Retrieval* 1(3), 233–334. DOI: http://doi.org/10.1561/1500000005.

Koolen, Corina, Karina van Dalen-Oskam, Andreas van Cranenburgh, and **Erica Nagelhout.** 2020. “Literary Quality in the Eye of the Dutch Reader: The National Reader Survey.” *Poetics* 79: 101439. DOI: https://doi.org/10.1016/j.poetic.2020.101439.

Lovinescu, Eugen. (c. 1928) 1981. *Istoria literaturii române contemporane*. Vol. 3. Reprint. Bucharest: Minerva.

Mendenhall, Thomas Corwin. 1887. “The Characteristic Curves of Composition.” *Science* 9(214): 237–249. DOI: https://doi.org/10.1126/science.ns-9.214S.237.

Moretti, Franco. 2005. *Graphs, Maps, Trees*. London: Verso.

———. 2013. *Distant Reading*. London: Verso.

———. 2017. “Patterns and Interpretation.” *Stanford Literary Lab Pamphlet* 15. September. Accessed November 10, 2022. https://litlab.stanford.edu/LiteraryLabPamphlet15.pdf.

Mosteller, Fredrick, and **David L. Wallace.** 1964. *Inference and Disputed Authorship: The Federalist. Addison-Wesley Series in Behavioral Science: Quantitative Methods*. Reading, MA: Addison-Wesley Publishing Company, Inc.

Murray, Simone. 2019. “Authorship.” In *The Oxford Handbook of Publishing*, edited by Angus Phillips and Michael Bhaskar, 38–54. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/oxfordhb/9780198794202.013.17.

North, Joseph. 2013. “What’s ‘New Critical’ about ‘Close Reading’? I. A. Richards and His New Critical Reception.” *New Literary History* 44: 141–157. https://www.jstor.org/stable/24542542. DOI: https://doi.org/10.1353/nlh.2013.0002.

Primorac, Antonija. 2023. “Sherlock Holmes and His Doppelgänger: For an Anti-Atlas of World Literature.” In *Anti-Atlas: Towards a Critical Area Studies*, edited by Tim Beasley-Murray, Wendy Bracewell, and Michał Murawski. London: UCL Press (FRINGE Series).

Radway, Janice A. (1984) 1987. *Reading the Romance: Women, Patriarchy, and Popular Literature*. Chapel Hill: University of North Carolina Press. Reprint. London: Verso.

Richards, Ivor Armstrong. (1924) 2017. *Principles of Literary Criticism*. London: Routledge & Kegan Paul. Reprint. New York: Routledge.

———. (1929) 1978. *Practical Criticism: A Study of Literary Judgment*. New York: Harcourt, Brace and Company. Reprint. London: Kegan Paul, Trench, Trubner & Co.

Riddell, Allen Beye. 2014. “How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models.” In *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, edited by Matt Erlin and Lynne Tatlock, 91–114. Rochester, NY: Camden House.

Rose, Mark. 1993. *Authors and Owners: The Invention of Copyright*. Cambridge, MA: Harvard University Press.

Sari, Yunita, Mark Stevenson, and **Andreas Vlachos.** 2018. “Topic or Style? Exploring the Most Useful Features for Authorship Attribution.” *Proceedings of the 27th International Conference on Computational Linguistics*, edited by Emily M. Bender, Leon Derczynski, and Pierre Isabelle, 343–353. Santa Fe: Association for Computational Linguistics.

Schaeffer, Jean-Marie. (1989) 2006. *¿Qué es un género literario?* Translated by Nicolás Campos Plaza and Juan Bravo Castillo. Madrid: Akal.

Schöch, Christof. 2017. "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama." *Digital Humanities Quarterly* 11(2). Accessed November 10, 2022. http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.

Schreibman, Susan, Ray Siemens, and **John Unsworth,** eds. 2015. *A New Companion to Digital Humanities*. Malden, MA: John Wiley & Sons. DOI: https://doi.org/10.1002/9781118680605.

Sowinski, Bernhard. 2013. "Stil." In *Historisches Wörterbuch der Rhetorik Online*, edited by Gert Ueding. Berlin: De Gruyter. DOI: https://doi.org/10.1515/hwro.8.stil.

Spitzer, Leo. (1928) 1931. *Stilstudien I: Sprachstile*. Reprint. München: Max Hueber.

Stamatatos, Efstathios. 2009. "A Survey of Modern Authorship Attribution Methods." *Journal of the American Society for Information Science and Technology* 60(3): 538–556. DOI: https://doi.org/10.1002/asi.21001.

Todorov, Tzvetan. (1978) 1990. *Genres in Discourse*. Translated by Catherine Porter. Cambridge, MA: Cambridge University Press.

Tomașevski, Boris. (1925) 1973. *Teoria literaturii: Poetica*. Translated by Leonida Teodorescu. Bucharest: Univers.

Underwood, Ted. 2014. *Understanding Genre in a Collection of a Million Volumes, Interim Report*. Figshare. DOI: https://doi.org/10.6084/m9.figshare.1281251.

———. 2017. "A Genealogy of Distant Reading." *Digital Humanities Quarterly* 11(2). Accessed November 10, 2022. http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html.

Van Dalen-Oskam, Karina. 2023. *The Riddle of Literary Quality*. Amsterdam: Amsterdam University Press. DOI: https://doi.org/10.5117/9789048558148_ch07.

Vickers, Brian. 2011. "Shakespeare and Authorship Studies in the Twenty-First Century." *Shakespeare Quarterly* 62(1): 106–142. http://www.jstor.org/stable/23025619. DOI: https://doi.org/10.1353/shq.2011.0004.

Wellek, René, and **Austin Warren.** (1942) 1956. *Theory of Literature*. Reprint. New York: Harcourt, Brace and Company.

———. (1942) 1967. *Teoria literaturii*. Translated by Rodica Tiniș. Bucharest: Editura pentru literatură universală.

Yule, George Udny. 1939. "On Sentence-Length as a Statistical Characteristic of Style in Prose: With Application to Two Cases of Disputed Authorship." *Biometrika* 30(3/4), 363–390. DOI: https://doi.org/10.2307/2332655.