Chapter 1

Introduction

1.1. Social and cultural implications of new media

The digital medium has opened a range of innovative possibilities for capturing, analysing and sharing information, and, as is the case in virtually all academic disciplines, a growing number of scholars in the field of literary studies are currently trying to harness the countless affordances of computational methods. As has been shown in numerous studies, digital technologies can radically transform the ways in which we can examine works of literature. Based on their capacity to store large amounts of data, and on the concomitant ability to perform extensive and complicated calculations within milliseconds, computers can help scholars to discover remarkable trends or correlations within massive collections of texts. Such analyses often reveal patterns which we could never see before, taking place on a scale that far exceeds the human bandwidth. Many of these technical possibilities have already been seized productively by scholars in the field of literary history. Digital methods can be used, for instance, to study the stylistic differences between texts from different historical eras, or to trace the rise and fall of literary genres in their entirety. Despite the many technological advances, it can be observed that, within the field of literary studies at large, the use of computers is still limited. The experiments with computer-assisted forms of reading have largely remained confined to a small group of pioneers. This slim uptake can partly be explained through the fact that it is still unclear whether computational and quantitative methods can genuinely expand or transform the ways in which we can interpret literary works. Such a capacity to provoke and to enhance hermeneutic interactions with texts is of crucial importance, since literary scholarship typically focuses on the quality and the meaning of literary texts, and on the various and intricate ways in which this meaning is produced. The main objective of this thesis is to understand the possibilities and the limitations of a computer-based literary criticism. Additionally, this study aims to illuminate the various ways in which the literary criticism that can be enabled by computational methods differs from the more traditional forms of criticism, which are generally based on detailed analyses of paper-based texts.

Because of this focus on the differences between computer-assisted scholarship on the one hand and scholarship based on printed resources on the other, this study fits within a longer tradition of research into the social and cultural implications of new media. This study is centrally based on the assumption that scholarship does not function autonomously, and that the methodologies and the

epistemology of academic disciplines are informed, partly at least, by the material properties of the media that are used to capture or to disseminate results. Various theorists have claimed, more broadly, that the technologies that we use to communicate are rarely neutral, and that they often influence us in ways that escape our awareness. Although this thesis mainly concentrates on a new, screen-based era of our textual history, this pivotal question about the cultural and psychological implications of new media also occupied many of the scholars who have studied earlier stages of the history of textual transmission. In the seminal work *Orality and Literacy*, for instance, Walter Ong focuses on the introduction of writing among previously oral cultures, referring to this development as "the technologising of the word". As is the case for more recent technologies such as the telephone or the tablet computer, the written word is a tool, or a "manufactured product"[1] through which human beings can expand and enrich their natural capacities for communication. In literate cultures, authors can externalise their thought processes, and once words have been consolidated on an inscription medium, the message assumes characteristics which are absent in the case of unmediated interaction. An obvious innovation was that the written message can be transferred to other locations, and that it could be preserved over time. In *The World on Paper*, David Olson argues, moreover, that the development of writing has had profound cognitive and psychological effects, and that the adoption of this technology directly shaped our "modern conception of the world and our modern conception of ourselves".[2] Written words, importantly, can be edited and corrected, and the possibility to reflect on particular phrases in turn stimulated a greater accuracy of formulation. Since a text that is recorded is also separated physically from its author, these impersonal texts can be scrutinised critically and objectively by other readers. By virtue of features such as these, it is alleged that the introduction of writing and reading decisively fostered man's capacity to think rationally and analytically.[3]

According to Ong, the advent of the printed book in the early modern period, and the introduction of the computer in the second half of the twentieth century ought to be viewed primarily as stages within a larger process in which the machinery to support reading and writing grew increasingly more sophisticated. Manuscript writing, in fact, "initiated what print and computers only continue", namely, "the reduction of dynamic sound to quiescent space".[4] While there is degree of continuity, it can also be observed that the introduction of new tech-

[1] Walter Ong, *Orality and Literacy: The Technologizing of the Word* (London: Routledge 2002), p. 78.
[2] David Olson, *The World on Paper: The Conceptual and Cognitive Implications of Writing and Reading* (Cambridge: Cambridge University Press 1994), p. 282.
[3] Ong emphasises that "abstractly sequential, classificatory, explanatory examination of phenomena or of stated truths is impossible without writing and reading". See Walter Ong, *Orality and Literacy: The Technologizing of the Word*, p. 8.
[4] Ibid., p. 81.

nologies often coincides with a range of transformative social and cultural effects. The manifold changes that followed the advent of the printed book form a case in point. In her influential study *The Printing Press as an Agent of Change*, Eisenstein argued that the invention of movable type in the second half of the 15th century, and the widespread availability of academic texts that it enabled, formed the driving force behind the unprecedented scientific advances in the early modern period. The printing press greatly extended the range and the variety of "the reading matter that was being surveyed at one time by a single pair of eyes".[5] Eisenstein argues that the impact of the printing press was based, to a large extent, on the capacity to make large numbers of texts available in a fixed and a stable form. The availability of a fixed text enabled researchers at different geographic locations to discuss text fragments which were exactly identical, and it encouraged scholars to build cumulatively on earlier ideas and discoveries.

The details of the transformative processes that are described in Eisenstein's study have frequently been contested, however. The claim that the technology of print inescapably fostered a standardisation and a systematisation of scientific knowledge has been challenged, for instance, in Adrian Johns' monograph *The Nature of the Book*. Johns stresses that fixity is not an inherent characteristic of print, but, rather, a quality which is transitive and historically contingent.[6] Printed texts have assumed a degree of fixity only as a result of the fact that particular agents have deliberately nurtured this aspect. Johns illustrates the constructivist argument by explaining that the fundamental instability that resulted from the many cases of plagiarism and piracy in the English publishing industry in the sixteenth century crucially undermined the trustworthiness of texts. Since the natural sciences depended acutely on accurate representations of scientific data, institutions such as the Royal Society purposely developed systems for the registration of authoritative works and for the protection of authorship.[7] Whereas Eisenstein postulates that the products of the printing press initiated profound social and cultural changes, Johns argues, inversely, that the features of print were shaped decisively by social forces. Harvey Graff stresses, likewise, that the printed codex is not inherently an agent of change, and that its impact is "determined by the manner in which human agency exploits them in a specific setting".[8]

The debate between Eisenstein and Johns concentrates, to a large extent, on the causality in the relationship between the printing press and the broader social

---

[5] Elizabeth Eisenstein, *The Printing Press as an Agent of Change: Communications and Cultural Transformations in Early Modern Europe* (Cambridge, New York: Cambridge University Press 1979), p. 289.

[6] Adrian Johns, *The Nature of the Book: Print and Knowledge in the Making* (Chicago: University of Chicago Press 1998), p. 19.

[7] Ibid., p. 475, and passim.

[8] Harvey J. Graff, *The Labyrinths of Literacy: Reflections on Literacy Past and Present* (London: The Falmer Press 1987), p. 19.

context in which these technological innovations are adopted. As such, the debate can be linked to a broader and more fundamental debate about the question whether or not technology can autonomously determine history. The technological determinism theory, at one extreme end of the spectrum, posits that "technological developments take place outside society, independently of social, economic, and political forces" and that "that technological change causes or determines social change".[9] Strict versions of the theory view the rise of new technologies as an exogenous process and emphasise that users of the new tools and devices have no choice but to accept the changes that are imposed. Critics of technological determinism have drawn attention, additionally, to the crucial importance of the social and political environment in which the technologies are developed and implemented. An extreme repudiation of the autonomous impact of technology may lead, nevertheless, to a form of social determinism, in which the consequences of new technologies are viewed exclusively as an outcome of social and political processes.

Most recent theories on the relation between technology and history take a stance which is located judiciously in between the two forms of determinism. While the details of the conjectures on the societal consequences of technologies differ, most theorists agree that technologies do not initiate changes autonomously, and that there is often a complicated interconnection between the nature and the impact of technological innovations and their broader social context. The set of theories which Keith Grint refers to collectively as the socio-technical systems approach claims that users generally have the freedom to use technologies in particular ways, and that, as a consequence, technology does not inexorably lead to particular pre-defined results. Authors who follow this approach concede "varying degrees of consequence to technology and social forces in a pluralistic net".[10] Contrary to what is claimed by technological determinism, the impact of technology may vary along with cultural, political and economic differences. Thomas Hughes stresses that tools and devices need to be studied as components in more encompassing "technological systems", which, next to the "physical artefacts" themselves, also encapsulate "organisations", scientific documentation, "legislative artefacts" and "natural resources".[11] By redefining technologies as much broader aggregates, comprising many different agents and artefacts, Hughes essentially dissolves the dialectic between technology and society. The assumption that technological developments take place outside of history has been contested nota-

[9] Sally Wyatt, "Technological Determinism Is Dead: Long Live Technological Determinism", in: *The Handbook of Science & Technology Studies*, (Cambridge: MIT Press 2008), p. 168.

[10] Keith Grint, *The Machine at Work: Technology, Work, and Organization* (Cambridge: Polity Press 1997), p. 12.

[11] Thomas Hughes, "The Evolution of Large Technological Systems", in: Wiebe Bijker, Thomas Hughes, & Trevor Pinch (eds.), *The Social Construction of Technological Systems: New Direxions in the Sociology and History of Technology*, (Cambridge: MIT Press 1987), p. 51.

bly by Langdon Winner, who has conceptualised technologies as "political phenomena in their own right".[12] Tools and devices are typically developed within a particular social context, and these tools consequently reflect the proclivities and, in some cases, the political beliefs of the original inventors. Next to serving their publicly stated goals, technological developments often serve to "enhance the power, authority, and privilege of some over others".[13]

The various theories on the relationship between history and technology provide a useful background for the analysis of the cultural and social changes that can follow the introduction of new technologies for the transmission of knowledge, of which the printed codex is one example. Two central principles may be proposed. It seems reasonable to assume, first, that the concrete features of media technologies are not given a priori, and that they are still pliable, within certain boundaries. Media have particular technological possibilities and limitations, which subsequently imply consequences for a wide range of aspects, including the number of modalities that can be disseminated, the level of interactivity, the speed with which messages can be distributed and the potential fixity of messages. The material properties and the technological possibilities of media set a range of options which may potentially be exploited. In many cases, these qualities also encourage or favour specific types of uses.[14] Second, the question whether or not these possibilities are enacted depends on the needs and the resourcefulness of the human agents who appropriate the technology. The implementation of technologies takes place within a socio-technical environment, and users of these tools need to acknowledge the relevance and the utility of the functions that are offered. These two principles can clarify the differences between Eisenstein's and Johns' arguments. According to Johns, Eisenstein's study places the development of print "outside history"[15] and considers fixity to be an inherent feature of print. Johns, to the contrary, posits a constructivist approach in which the features of print are governed by social and cultural factors. The printed medium can potentially be used to disseminate fixed texts, but this feature may also be defective under different historical or cultural circumstances. Printed texts can be made stable after particular communities of users have acknowledged the desirability of such stability.

Following the principles that have been outlined above, the implications of the digital medium can be analysed by considering the technological possibilities and limitations that follow from the basic material properties. Specific material properties have implications for the ways in which messages can be produced,

---

[12] Langdon Winner, "Do Artefacts Have Politics?", in: *Daedalus*, 190:1. Modern Technology: Problem or Opportunity? (1980), p. 123.

[13] Ibid., p. 125.

[14] Adriaan van der Weel, "Pandora's Box of Text Technology", in: *Jaarboek Voor Nederlandse Boekgeschiedenis*, (Nijmegen: Vantilt 2013).

[15] Adrian Johns, *The Nature of the Book: Print and Knowledge in the Making*, p. 19.

distributed and consumed. These properties ought to be studied in relation to the question whether or not these properties are recognised and exploited within particular communities. The digital medium derives many of its crucial qualities from the fact that it is possible to produce and to disseminate text with great ease and at an unparalleled speed. In a sense, this capacity may be viewed as a continuation of a development which quickened after the development of the mechanical printing press. Adriaan van der Weel emphasises that the process of printing was designed to increase the speed with which titles could be made available, and, related to this, to raise the number of copies that could be produced. Whereas the printing press could in theory be used to produce a low number of copies, the investments needed to finance the labour-intensive preparations of a work could be recouped only by selling many books. Since this economical imperative forced publishers to secure large print runs, it can be observed that the inherent properties of print strongly favoured particular types of usage over other applications.[16] This growth in resources enabled scholars to build cumulatively on ideas and discoveries that had been recorded previously, and to amalgamate and synthesise these in order to produce new texts. Innovations in the technology of printing, developed and implemented by engineers such as Lord Stanhope and Friedrich Koening in the nineteenth century,[17] eventually led to the mass production of books, and this overabundance of publications in turn induced many contemporary readers to complain about the sense of information overload.

This process of proliferation further intensified by several orders of magnitude, nonetheless, on today's worldwide web. Digital texts are essentially non-material entities, and, as a result of this, many of the practical challenges posed by the distribution of paper-based publications no longer apply. In addition, while the channels for the distribution of information were previously monopolised by professional publishers, these are now within the reach of virtually anyone with an internet connection. Since online publication is often recognised as a means to enhance scholarly impact, the opportunity to disseminate scholarly content quickly and without obstacles has been seized by many scholars. Texts and data can be made available through repositories, weblogs or on wikis, and, in this way, scholars can engage directly with their peers. Information is frequently made accessible free of charge and free of copyright and licensing restrictions, and this clearly has consequences for the dissemination of information. Richard Lanham emphasises

[16] Adriaan van der Weel, *Changing Our Textual Minds : Towards a Digital Order of Knowledge* (Manchester: Manchester University Press 2011), p. 82. Febvre and Martin explain similarly that the printed book's capacity to act as a "force of change" can be connected, to a large extent to the increased speed of copying and the general growth in the number of titles. See Lucien Febvre & Henri-Jean Martin, *The Coming of the Book: The Impact of Printing 1450-1800* (London: NLB 1976), p. 249ff.

[17] Asa Briggs, *A Social History of the Media: From Gutenberg to the Internet* (Cambridge: Polity 2002), p. 2.

that, whereas "the codex book limits the wisdom of Great Books to students who are Great Readers",[18] the digital medium has extended the access to scientific resources to audiences beyond the direct scholarly community. Peter Shillingburg notes, in a similar vein, that "[w]hat Gutenberg did to democratize books and other texts, the World Wide Web has done to democratize information".[19]

The principle that the characteristics of media offer a range of technological possibilities, and that its features are malleable, within limits, is evinced by the various attempts to endow digital texts with a degree of fixity and authority. On the web, there are no natural authorities who can monitor who publishes information, and, equally crucially, who removes information.[20] This has the effect that online resources generally lack stability. Eisenstein emphasises that the 'scientific revolution' in the early modern period was stimulated strongly by the fact that large numbers of readers had access to stable texts that "provided a common base for later disputes among scholars".[21] Since science and scholarship typically aim to produce durable and authoritative knowledge, academic publishers and university libraries have tried to develop mechanisms to address the shortcomings associated with the ethereality of digital documents. Measures include the assignment of persistent identifiers for publications and for authors, the stimulation of use of typographically stable formats such as PDF, and the development of technical solutions in the field of digital rights management and long-term preservation.

1.2. Implications for scholarship

Andy Clarke claims that when we use analogue or digital technologies such as smartphones, notebooks or calculators in order to think and to produce new knowledge, such machinery ought to be viewed as extensions of the mind. These technologies have the consequence that particular cognitive processes can be

[18] Richard Lanham, *The Electronic Word: Democracy, Technology, and the Arts* (Chicago: University of Chicago Press 1993), p. 39.

[19] Peter Shillingsburg, *From Gutenberg to Google: Electronic Representations of Literary Texts* (Cambridge: Cambridge University Press 2006), p. 2.

[20] It is often difficult, nonetheless, to fully delete information from the web. While web pages or files can clearly be removed from a web server, search engines which have crawled the site may continue to supply snippets of the text. Additionally, other web sites may have copied information, without the knowledge of the original source of this information. A study that was conducted by Hennessey and Ge has demonstrated, nevertheless, that the majority of web resources which were referenced in scientific articles could no longer be accessed after a period of ten years. The authors found that the median lifespan of web pages was only 9.3 years and that a mere 62% of the web resources which were referenced had actually been archived. See Jason Hennessey & Steven Ge, "A Cross Disciplinary Study of Link Decay and the Effectiveness of Mitigation Techniques.", in: *BMC bioinformatics*, 14 Suppl 1:14 (9 January 2013).

[21] Elizabeth Eisenstein, *The Printing Press as an Agent of Change: Communications and Cultural Transformations in Early Modern Europe*, p. 350.

realised "by structures and processes located outside the human head".[22] Clark surmises that the separation between the human mind and its external environment is arbitrary, as physical objects can perform functions for the same purposes as processes within the human brain.[23] It step with Clark's extended mind theory, it can be argued that any changes in the technologies that are used to support and to stimulate cognitive process have fundamental repercussions, not only on the manner in which the results of academic enquiry can be disseminated, but also on the manner in which research can be conducted.

This thesis concentrates, for an important part, on the ways in which digital technologies are transforming scholarship. At present, we are witnessing a transition from a system in which scholarly knowledge is disseminated predominantly via paper-based media to a situation in which these analogue forms of output are increasingly supplanted or supplemented by digital forms of scholarly output. Particularly in the natural sciences and in the life sciences, existing practices have been transformed immensely by the numerous new possibilities in the field of network computing and information technology. The type of research that is enabled through innovations in ICT is often referred to as "e-Science". It is commonly viewed as a confluence of three technological developments.[24] The first of these is the unprecented growth of the availability of research data. In recent years, the phrase "big data" has been used recurrently to denote the ever growing volumes of data that some research projects or commercial enterprises are facing. Current research programmes, especially in fields such as high-energy physics, astronomy and genomics, often use digital measuring devices that spawn quantities of machine-readable data at rates which outstrip the possibilities to analyse them. Bell et al. note that "some areas of science are facing hundred- to thousandfold increases in data volumes from satellites, telescopes, high throughput instruments, sensor networks, accelerators, and supercomputers, compared to the volumes generated only a decade ago".[25] As a result, it is often difficult for researchers to study the data about these phenomena directly. The only way in which

[22] Andy Clark, *Supersizing the Mind: Embodiment, Action, and Cognitive Extension* (Oxford: Oxford University Press 2008), p. 76.

[23] To clarify his argument, Clark cites an exchange between physicist Richard Feynman and historian Charles Weiner. When Weiner remarked that Feynman's notes and sketches on paper represented a record a Feynman's work, Feynman retorted that the archive is not record of his work and that writing notes on paper must be viewed as working in itself. According to Clark, this "loop into the external medium was integral to his intellectual activity" to such an extent that "Feynman was actually thinking on the paper". See ibid., p. xxv.

[24] Anne Beaulieu & Paul Wouters, "E-Research as Intervention — E-Research: Transformation in Scholarly Practice", in: Nicholas Jankowski (ed.), *E-Research: Transformation in Scholarly Practice*, (London: Routledge), p. 55.

[25] Gordon Bell, Tony Hey & Alex Szalay, "Beyond the Data Deluge", in: *Science*, 323:5919 (6 March 2009), p. 1297.

researchers can cope with such vast data collections is by letting data analysis software produce summaries or abstractions of these data sets.

Various authors have argued that the staggering rise in the quantity of data may stimulate, or perhaps even necessitate, a new form of research. In an influential lecture delivered to the American National Research Council in 2007, computer scientist Jim Grey argued that the move to a more data-intensive science, which uses groundbreaking technologies for the analysis and the visualisation of these data, may legitimately be viewed as a paradigm shift. In a traditional setting, scientists firstly formed hypotheses, based on an explanatory theory, before they conducted experiments to corroborate or refute these hypotheses. In data-driven research, computers initially search for patterns or for regularities in the data, allowing researchers to search for hypotheses that may explain these statistical phenomena in retrospect.[26] In very a similar vein, Chris Anderson, in his article "The End of Theory", argues that when research data are available on the petabyte scale, this "forces us to view data mathematically first and establish a context for it later".[27] When vast datasets are combined with statistical algorithms and applied mathematics, such advanced number-crunching techniques largely supersede the need to formulate explanatory theories. In such data-intensive fields, scientists increasingly rely on sophisticated search tools and visualisation techniques which enable them to trace patterns and to make new discoveries on the basis of massive sets of research data. A new type of methodology thus appears to be emerging, in which discoveries are mainly made by mining existing data sets.

Next to an intensification of the use of digital data, e-Science also entails a growing reliance on grid-computing facilities and networks which can ensure that these collections of data can be analysed at locations other than the sites on which these data originated. Hey and Trefethen write that the "two key technological drivers of the IT revolution are Moore's Law - the exponential increase in computing power and solid-state memory - and the dramatic increase in communication bandwidth made possible by optical fibre networks using optical amplifiers and wave division multiplexing".[28] Grid computing means that researchers do not only exchange data, but that they also share computing resources to manage and to process these data. A third development which is generally considered to be part of e-Science is the notion that academic studies tend to become more collaborative. The impetus to cooperate is usually connected to a growing specialisation and an

[26] Jim Gray, "Jim Gray on eScience: A Transformed Scientific Method", in: Tony Hey, Stewart Tansley, & Kristin Tolle (eds.), *The Fourth Paradigm: Data-Intensive Scientific Discovery*, (Redmond: Microsoft Research 2009), p. xix.

[27] Chris Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", in: *Wired Magazine*, 16:07 (2008), n.pag.

[28] Tony Hey & Ann Trefethen, "The Data Deluge: An E-Science Perspective", in: Fran Berman, Geoffrey Fox, & Tony Hey (eds.), *Grid Computing: Making The Global Infrastructure a Reality*, Chichester: Wiley 2003, p. 810.

increasing complexity of scientific problems. Many questions can only be answered adequately if expertise from different disciplines can be combined and if the amount of work that needs to be carried out can be divided over different individuals. By virtue of the internet and grid computing facilities, geographically dispersed teams of researchers can share some of their data sets, and carry out analyses on the accumulated resources. To support such collaborative work, teams of researchers often make use of online environments in which they can share relevant data and research tools. To describe such collaborative environments, terms such as "collaboratories"[29] or "Virtual Research Environments"[30] are often used.

Reports and articles that outline the advantages that may emanate from an upsurge in the use of technology often display a remarkable optimism.[31] Such hopefulness and blatant exuberance already pervaded Douglas Engelbart's 1962 essay "Augmenting Human Intellect". Engelbart argued that machines can bring "better comprehension, the possibility of gaining a useful degree of comprehension in a situation that previously was too complex".[32] Numerous texts about the nature of e-Research expound the positivist belief that when more digital data are made available, and when computers become faster, this will ultimately lead to an increase in the number of scientific discoveries.[33] John Wilbanks, for instance, writes that "[d]ata-intensive science, if done right, will mean more paradigm shifts of scientific theory, happening faster, because we can rapidly assess our worldview against the 'objective reality' we can so powerfully measure".[34] A comparable belief

---

[29] William Wulf, "The Collaboratory Opportunity", in: *Science*, 261:5123 (1993), p. 854.

[30] Annamaria Carusi & Torsten Reimer, *VRE Collaborative Landscape Study*, (London: 2010).

[31] Sally Wyatt notes that authors who have defended the technological determinism theory likewise subscribed to the simplistic notion that "technological progress equals social progress". See Sally Wyatt, "Technological Determinism Is Dead: Long Live Technological Determinism", p. 168.

[32] Douglas Engelbart, *Augmenting Human Intellect: A Conceptual Framework* (Menlo Park Calif.: Stanford Research Institute 1962), p. 1.

[33] It must be added that there are also many scholars who have denounced the consequences of technological advances. Books such as Giedion's *Mechanization Takes Command*, Lewis Mumford's *Technics and Civilization*, and Jacques Ellul's *The Technological Society* mainly stress the unfavourable effects. Mumford and Ellul both argue that technology creates a threatening environment in which human beings are enslaved by the working methods that are imposed by artificial devices. Ellul emphasises the dehumanising effects of technological systems, which demand efficiency and rationality within all the domains they are applied to, and which "bring mechanisation to bear on everything that is spontaneous and irrational" (pp. 78-79). See Sigfried Giedion, *Mechanization Takes Command: A Contribution to Anonymous History. (Oxford University Press 1948),* Lewis Mumford, *Technics and Civilization* (New York: Harcourt Brace and Co. 1934) and Jacques Ellul, *The Technological Society* (New York: Alfred A. Knopf 1973), pp. 78–79. In his foreword to the 1973 edition, Robert Merton notes that Ellul emphasises "the erosion of moral values bought about by technicism" (p. v).

[34] John Wilbanks, "I Have Seen the Paradigm Shift, and It Is Us", in: Tony Hey, Stewart Tansley, & Kristin Tolle (eds.), *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Redmond: Microsoft Research 2009, p. 210.

in the beneficial effects of eResearch can be found in Borgman, who explains that "[d]ata have become an important form of research capital, enabling new questions to be asked", and that "[t]ext and data mining promise everything from drug discovery to cultural enlightenment".[35]

Many attempts to stimulate the access to data and publications are similarly based on the assumption that, when scholarly resources are made available as widely as possible, this functions as a catalyst for the generation of new knowledge. The *OECD Declaration on Access to Research Data from Public Funding*, for instance, highlights the importance of making underlying research data available for reuse beyond the research project in which they were initially produced.[36] If data can be shared among colleagues who are working on similar questions, these related studies can reduce their data collection efforts, and move more quickly to the discovery phase. It is alleged that new research projects will have access to increasingly large quantities of data, which means that the scope of these studies can also be extended accordingly. New studies may also exploit the data in ways that were not envisaged when they were originally created. Furthermore, it is also maintained that continued access to research data will improve the transparency of the research process. When the data that underpins a specific study are shared, this enables peers to replicate and to verify the claims that are made by that study. Through such forms of openness, cases of incorrect reasoning, or, worse, of deliberately misreported or fraudulent data may eventually be identified and exposed more efficiently. Borgman confirms that that, "[i]f the data are available, then a more rigorous review of the scholarship becomes possible".[37]

The concrete ways in which technologies are implemented are often contingent on the ability of adopters to recognise their utility. As was discussed above, e-Science entails data-intensity, grid computing, and an intensification of collaboration. These three components are not equally relevant for all academic fields, however. While a growing number of disciplines rely on the use of new media and of communication networks, only a few of them actually demand distributed high-performance computing facilities. Beaulieu and Wouters explain that the term "e-Science" focuses specifically on collaborative, data-intensive and grid-enabled research projects, and that "e-Research" is a more inclusive term, which refers more broadly to the various ways in which computer-based methodologies can transform scholarly and scientific practices.[38] Furthermore, while the scope of term "e-Science" is usually reserved for studies in the natural sciences, "e-

[35] Christine Borgman, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet* (Cambridge: MIT Press 2007), pp. xvii–xviii.

[36] *OECD Principles and Guidelines for Access to Research Data from Public Funding*, (Paris: 2007).

[37] Christine Borgman, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*, p. 127.

[38] Anne Beaulieu & Paul Wouters, "E-Research as Intervention — E-Research: Transformation in Scholarly Practice", p. 55.

Research" also encompasses other disciplines, such as the social sciences and the humanities.[39] The impact of digital data is growing generally, but this development has not affected all scholarly field uniformly. Making a shift to e-Research, in most cases, demands "an active transformation, in which the models of research prevalent in the relevant field shape the way ICT is conceptualized and used".[40]

While terms such as e-Research and e-Science are inherently broad and interdisciplinary, there are also numerous attempts to explore the impact and affordances of digital technology within specific academic fields. In many existing disciplines, new subfields have been minted which are frequently referred to using a coinage that fuses the standard designation of the discipline with the term *informatics*. This latter term, in turn, has been defined as "the science and practice dealing with the effective collection, storage, retrieval and use of information".[41] *Construction informatics*, for example, has been defined as an "interdisciplinary discipline filling the gap between computer science and construction".[42] One of the most significant examples of the impact of the use of big data collections on traditional practices can be found in the field of biology. *Bioinformatics* is "the interdisciplinary toolset for applying computer science, mathematics, and statistics to the classification and analysis of biological information". While it is generally agreed that bioinformatics helps researchers to address traditional questions, the scale and the speed at which data-driven research can operate often create the possibility to arrive at new types of answers. As evidenced by the results of the *Human Genome Project*, insights from informatics can "result in unprecedented power" and can help biologists to "handle large quantities of data and probe the complex dynamics observed in nature".[43]

1.3. Digital humanities

Computational techniques have also affected research that takes place within the humanities, and the scholarly area which investigates the symbiosis of informatics and humanities research is mostly known as the digital humanities. The objectives of the digital humanities are twofold. First, the field focuses on the various ways in which the computer can be used to investigate traditional questions in the humanities. Second, the field also studies the phenomenon of computation from a

[39] Marc Wilhelm Küster, Thomas Selig & Julianne Nyhan, *Report on eHumanities: research topics relevant in the Computer Science*, (2010), p. 7.

[40] Anne Beaulieu & Paul Wouters, "E-Research as Intervention — E-Research: Transformation in Scholarly Practice", p. 55.

[41] T. Saracevic, "Information Science", in: M. J. Bates (ed.), *Encyclopedia of Library and Information Sciences*, 3rd editio, (New York: Taylor and Francis), p. 2570.

[42] Žiga Turk et al., *ICT Ontological Framework and Classification*, (Ljubljana: 2002), p. 5.

[43] N. M. Luscombe, D. Greenbaum & M. Gerstein, "What Is Bioinformatics? A Proposed Definition and Overview of the Field", p. 346.

humanities perspective, and aims to understand the epistemological and the methodological implications of using computers in humanities research.[44] Perhaps to a larger extent than in other fields, the application of the digital medium poses a number of challenges in disciplines that "illuminate the human record".[45]

Discipline-based assessments of the potential benefits of computation are necessary because different fields often adhere to unique methodological and epistemological traditions. Data-driven research sets a number of basic demands which may or may not be compatible with these traditions. First, a degree of consensus is needed as to what precisely constitutes data. Second, data-driven research requires a shared understanding of the methods that can be used to analyse these data. Researchers, third, need to share a common understanding of the overall rationale of these data analyses and of the manner in which the results of data analyses can contribute to the creation of new knowledge. This list of requirements is not exhaustive, but it seems evident that these requirements should minimally be met to ensure that the adoption of computational methods can be advantageous. A consensus on the nature and the purpose of data processing is most likely to be achieved in fields in the life sciences and in the natural sciences, which generally aim at producing objective and verifiable knowledge. Carl Hempel points out that an explanation is scientific if it makes a reference to a general and universally applicable law. General laws are "empirical generalizations connecting different observable aspects of the phenomena under scrutiny".[46] Since the ultimate objective of science is to understand these universal laws, scientists generally believe that a single correct explanation can be given for concrete events or phenomena, as each of these obey a single set of universal laws. Generally, this scientific approach also assumes that questions can be answered in a definitive and conclusive manner, and that analyses of data about these phenomena may help to provide these answers.

The conditions which are indispensable for a consequential application of digital methods are not necessarily present within humanities research. In many humanities fields, there is still some uncertainty as to what exactly constitutes

[44] This is a paraphrase of the definition provided in Kathleen Fitzpatrick, "The Humanities, Done Digitally", in: *Debates in the Digital Humanities*, University of Minnesota Press 2012. Fitzpatrick defines the digital humanities as "a nexus of fields within which scholars use computing technologies to investigate the kinds of questions that are traditional to the humanities, or, as is more true of my own work, ask traditional kinds of humanities-oriented questions about computing technologies" (p. 12)

[45] Susan Schreibman, Ray Siemens & John Unsworth, "The Digital Humanities and Humanities Computing: An Introduction", in: Susan Schreibman, Ray Siemens, & John Unsworth (eds.), *A Companion to Digital Humanities*, Oxford: Blackwell 2004, p. xxiii.

[46] Carl Hempel & Paul Oppenheim, "Two Models of Scientific Explanation", in: Yuri Balashov & Alexander Rosenberg (eds.), *Philosophy of Science: Contemporary Readings*, London and New York: Routledge 2002, p. 47.

data.[47] In reaction to the ubiquitous phrase "the data deluge", Anderson et al. suggest that, in relation to the humanities, it seems more apt to speak of a "complexity deluge", as the discipline deals with "a multiplicity of types of information, much of it highly dispersed, difficult to find and complex to use".[48] Furthermore, many of the benefits that are associated with e-Research demand that data can be captured in a structured and consistent format, and that there is agreement on how the data are to be analysed. Harvey points out that "[w]hile much (but far from all) data within the physical and biological sciences are relatively more comparable and can be deposited into common databases, no such 'common denominator' exists for social and humanistic data, since data types, sources, and collecting practices can vary so widely".[49]

A belief in universally applicable rules and laws is certainly not widespread among humanities researchers. Costis Dallas has shown that there are a number of marked epistemological differences between the humanities on the one hand and science, technology and medicine on the other. While the natural sciences are conventionally "experimental, dealing directly with the empirical domain viewed as a closed system", research in the humanities is "often hermeneutic, dealing with complex, agglomerative structures of argument manifested in the corpus of earlier scholarship".[50] When humanities scholars adopt digital research instruments, this simultaneously forces them to make a transition to an approach which is more similar to that of the natural sciences, and in which data and analytic procedures are more standardised. A reliance on empiricism and objectivity seems antithetical to many existing practices in the humanities, since, as noted by Salemans, humanities research is traditionally deductive rather than inductive. It starts with "the definition of subjective thoughts, ideas, hypotheses about the material or facts to be investigated". This deductive method often has negative connotations, and scholars who "do not want to be accused of subjective, and therefore unscientific research … feel obliged to replace their deductive research by inductive research".[51] Through the approbation of computational methods, scholars can move towards an

[47] See, for instance, Christine L. Borgman, "The Digital Future Is Now: A Call to Action for the Humanities", in: *Digital Humanities Quarterly*, 003:4 (2010).

[48] Sheila Anderson, Tobias Blanke & Stuart Dunn, "Methodological Commons: Arts and Humanities E-Science Fundamentals.", in: *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 368:1925 (2010), p. 3781.

[49] Diane Harley, *Assessing the future landscape of scholarly communication an exploration of faculty values and needs in seven disciplines*, (Berkeley, CA: Center for Studies in Higher Education 2010), p. 27.

[50] Costis Dallas, "Humanistic Research, Information Resources and Electronic Communication", in: J. Meadows & H. Boecker (eds.), *Electronic Communication and Research in Europe*, Luxemburg: 1998, p. 210.

[51] Ben Salemans, "The Remarkable Struggle of Textual Criticism and Text-Genealogy to Become Truly Scientific", in: Wido van Peursen, Ernst D. Thoutenhoofd, & Adriaan van der Weel (eds.), *Text Comparison and Digital Creativity: The Production of Presence and Meaning in Digital Text Scholarship*, Leiden, Boston : Brill 2010, pp. 113–114.

approach in which insights are derived from observable or quantifiable facts. In a similar vein, and with respect to philological and textual research, Van Peursen notes that "[t]he ability to sort, quantify, reproduce, and report text through computation would seem to facilitate the exploration of text as another type of quantitative data". The formalisation and the emphasis on explicitness can be considered "as a means to overcome the individualism and subjectivity that characterizes much philological research".[52] The aim is often interpretation and understanding. It is not incontrovertibly clear how understanding may ensue from data processing.

Within the humanities at large, there are a number of fields in which the empirical and inductive approach seem opportune. In fields such as archaeology and linguistics, there is often unanimity on the nature of research data and on the manner in which these resources ought to be analysed. As such, archaeology and linguistics clearly share a number of characteristics with the 'hard' sciences. In linguistics, research is typically based on the assumption that linguistic utterances follow a set of underlying principles and laws which may be exposed and described through sufficiently thorough analyses of empirical data. As in the natural sciences, a number of disciplines in the humanities aim at providing single answers to questions. Unsurprisingly, in archaeology and linguistics, the use of digital research instruments has also become fairly commonplace. Although computer-based research within these field continues to produce countless technical and organisational difficulties, linguists and archaeologists increasingly view computational tools as an integral part of their general methodology. The use of computational methods has advanced to such an extent that large international infrastructures needed to be built to disseminate and to curate tools and primary data for the international research community.[53] There are also a number of humanities disciplines, however, in which the application of the central concepts of e-Science seems less evident, such as philosophy or literary studies.

[52] Wido van Peursen, "Text Comparison and Digital Creativity", in: Wido van Peursen, Ernst D Thoutenhoofd, & Adriaan van der Weel (eds.), *Text Comparison and Digital Creativity : The Production of Presence and Meaning in Digital Text Scholarship*, Leiden, Boston: Brill 2010, pp. 12–13.

[53] The CLARIN project, for instance, aims to implement an international infrastructure for linguistic research by providing uniform access to the contents of distributed digital archives and to the various language and speech processing tools that have been developed. See <http://clarin.eu/>. At the same time, however, the accomplishments of the CLARIN project ought not to be overestimated. For many scholars, it is still difficult to effectively incorporate digital tools within their research processes, and the differences in methodologies still result in data sets which are ultimately difficult to reuse. See, amongst many other sources, *CLARIN-NL Annual Report 2014*, (2015).

1.4. Literary studies in the digital age

This thesis will focus on the ramifications of the introduction of digital technology within the academic field of literary studies. Baldick notes, in general terms, that literary criticism may include diverse activities such as "classification of a work according to its genre, interpretation of its meaning, analysis of its structure and style, judgement of its worth by comparison with other works, estimation of its likely effect on readers".[54] Literary critics aim to describe, to explain and to justify their subjective experience of a specific work or of a body of works. In *Principles of Literary Criticism*, I.A. Richards explains that scholars typically focus on questions such as "What gives the experience of reading a poem its value? How is this experience better than another?", and "How can experiences be compared? What is value?".[55] For a number of reasons, the application of digital tools seems inopportune in the field of literary studies. In most cases, hermeneutic practices are not standardised, and critics often use idiosyncratic methods for analyses of texts. Furthermore, unlike history or linguistics, literary studies is usually open-ended, and scholars do not aim to answer questions in a definitive way. The goal is generally to contribute to a specific debate, and not to end it. Critics who study Virginia Woolf "are not trying to solve Woolf", but they are trying to make sure that the discussion of Woolf's novels "continues into further and further reaches of intellectual depth".[56] The objective of a study is typically to produce a discourse in which the author tries to convince his peers of the validity of certain ideas. Insights about literary works change according to culture and over time, and the coexistence of multiple views and dissimilar interpretations is not necessarily viewed as problematic. George Steiner, in *Real Presences*, confirms that "[i]n aesthetic discourse, no interpretative-critical analysis, doctrine or programme is superseded, is erased, by any later construction". "Aristotle on mimesis and pathos", for instance, "is not superseded by Lessing or by Bergson", and the "Surrealist manifestos of Breton do not cancel out Pope's Essay on Criticism though they may well be antithetical to it".[57]

Whereas the field of literary studies has a number of characteristics which clearly complicate the adoption of computational techniques when attempting to answer the existing questions, a number of recent developments are likely to have important consequences for the manner in which literary texts can be investigated. The most notable of these follow from the vast increase in the number of texts that are available in a machine readable form. Numerous commercial and non-

[54] Chris Baldick, "Literary Criticism", in: Chris Baldick (ed.), *The Oxford Dictionary of Literary Terms*, Oxford: Oxford University Press 2008.

[55] Ivar Armstrong Richards, Principles of Literary Criticism. (New York: Harcourt Brace 1961), p. 2.

[56] Stephen Ramsey, "Algorithmic Criticism", in: Susan Schreibman & Ray Siemens (eds.), *A Companion to Digital Literary Studies*, Oxford: Blackwell 2008, p. 489.

[57] George Steiner, *Real Presences* (University Of Chicago Press 1991), p. 76.

commercial parties have decided to exploit the ease with which information can be disseminated on the web and have set up online repositories in which large volumes of digitised or born-digital texts can be made available to wider audiences. Important examples of such initiatives to extend and to improve access to textual materials include Project Gutenberg, the Open Content Alliance and the Million Book Project at Carnegie Mellon. Similarly, the collections Eighteenth Century Collections Online (ECCO) and Early English Books Online (EEBO) together offer researchers the possibility to search the contents of some 200,000 books published between the second half of the fifteenth century to the beginning of the nineteenth century. In some cases, texts are publicly available, but in other cases, a paid subscription is needed. Out of the many initiatives that have been launched to produce corpora of electronic texts, however, the most ambitious and most audacious programme is probably Google Books. At the 2004 Frankfurt Book Fair, Google first announced its plans to scan the holdings of libraries worldwide and to publish these scans together with the full text that was to be obtained through OCR. Various prestigious libraries participated in the project, including the University Library of Michigan, Harvard University Library, Stanford Green Library, The Bodleian Library at Oxford and New York Public Library. It is estimated that Google has currently digitised over seven million books.

Most of the projects that have been cited engage in mass-digitisation and aim to be as inclusive as possible by scanning complete book cases or even complete libraries. As is noted by Julia Flanders, for such projects, "storage is cheaper than decision making".[58] By contrast, there are also various examples of projects which are more limited in scope, and in which scholars have carefully prepared digital critical editions of the works of individual authors. The Rossetti Archive, for instance, is maintained at the University of Virginia under the editorship of Jerome McGann, and was developed to facilitate the scholarly investigation of the works of Dante Gabriel Rossetti.[59] On the project websites it is explained that all documents are encoded to allow for advanced searching. The *Algernon Charles Swinburne Project*, which was conducted at Indiana University, is very similar in scope, as it aims to provide "students and scholars with access to all available original works by Swinburne and selected contextual materials".[60] Next to these critical editions of the texts from a single author, there are also a number of scholarly textbases that focus on specific geographic areas or on specific genres. A first example is *CELT*, which was developed at University College Cork in order to "bring the wealth of Irish literary and historical culture [...] to the Internet in a rigorously scholarly and

[58] Julia Flanders, "The Productive Unease of 21st-Century Digital Scholarship", in: Melissa Terras, Julianne Nyhan, & Edward Vanhoutte (eds.), *Defining Digital Humanities*, Farnham: Ashgate 2013, p. 207.

[59] *The Rossetti Archive*, <http://www.rossettiarchive.org/> (18 March 2014)

[60] *The Algernon Charles Swinburne Project*, <http://swinburnearchive.indiana.edu/> (18 March 2014)

user-friendly project for the widest possible range of readers and researchers".[61] A second prominent example is the *Women Writers Project* at Brown University which is intended "to bring texts by pre-Victorian women writers out of the archive and make them accessible to a wide audience of teachers, students, scholars, and the general reader".[62]

Given the speed at which many digitisation projects proceed, the thought that, in the near future, all titles that have ever been published will be available in a digital format seems progressively less preposterous. The ease with which digital sources can be disseminated and accessed has already had enormous implications for the efficiency of scholarship. When rare and unique materials have been digitised, this often means that scholars can consult these materials on their screens, and that they can save themselves visits to remote libraries. Michael Hart, who founded Project Gutenberg, used the term "replicator technology" to describe to the idea that, once a work is available in a digital form, this text can be distributed among readers in an unlimited number of copies.[63] When the digital medium is used exclusively to optimise the process of providing access, however, this does not fundamentally alter the manner in which these materials are studied. Martin Mueller notes that digital archives such as EEBO and ECCO have primarily effectuated a "first-order increase in query potential".[64] The online availability of large collections of scans have made it easier for scholars to find relevant titles and to gain access to them. Once the titles have been located, however, scholars often print the files, and continue to study these texts in exactly the same way as they would study a codex book. There is regularly a disregard of the notion that digital resources also have a "second-order query potential",[65] in the sense that they can be restructured and queried in a manner that was previously impossible.

Electronic texts differ from texts on analogue media in a number of important ways. Whereas, in the case of paper-based publications, text and images are the only modalities that can be disseminated, a digital environment allows for the seamless convergence of various kinds of modalities, such as text, sound, images, audio and video. In addition, digital content is flexible and malleable. Walter Ong observed about printing that it "situates words in space" and that it "locks words into position".[66] The printing process casts the text in a "state of completion", and

---

[61] *CELT*, <http://www.ucc.ie/celt/> (18 March 2014)

[62] *Women Writers Project*, <http://www.northeastern.edu/nulab/women-writers-project-2/> (25 June 2013)

[63] Michael Hart, "The History and Philosophy of Project Gutenberg", 1992, <https://www.gutenberg.org/wiki/Gutenberg:The_History_and_Philosophy_of_Project_Gutenberg_by_Michael_Hart> (25 June 2013).

[64] Martin Mueller, "Digital Shakespeare, or towards a Literary Informatics", in: *Shakespeare*, 4:3 (September 2008), p. 288.

[65] Ibid.

[66] Walter Ong, *Orality and Literacy: The Technologizing of the Word*, p. 119.

texts consequently become "autonomous and indifferent to attack".[67] A digital text, by contrast, ultimately exists as a vast collection of bits within the computer's memory. Consequently, it is essentially a fluid and navigable entity which can be reshuffled or recomposed in support of specific scholarly interests. Applications can be developed that comb the text for particular fragments and patterns, or which combines strings from different contexts for the purpose of comparison. It is also feasible to isolate fractions of the text with specific properties and to perform calculations or other forms of manipulations on such excerpts.

In the introduction to his influential book *Principles of Literary Criticism*, I.A. Richards famously referred to the printed codex as "a machine to think with", and also compared the book to "a loom" on which authors can "re-weave some ravelled parts of our civilization".[68] Richards' metaphors underscore the notion that the book can be viewed as a feat of technology that simulates particular intellectual processes and that spurs the generation of new knowledge, through the convenient accessibility of recorded knowledge. Since scholars in the humanities currently have speedy and convenient access to more primary sources than were ever imaginable, and since the mechanisms with which scholars can search and retrieve these digital sources grow more and more sophisticated, it seems reasonable to expect that the digital machine can be generative of more encompassing or more diversified forms of thinking. It can be assumed, moreover, that methods and workflows that were developed originally for data-intensive projects in the natural sciences increasingly become relevant for literary studies.

The networked computer opens up a multitude of new possibilities for organizing, querying, visualising and disseminating texts, and, among a number of pioneering scholars, the potential of the digital medium has inspired a clear zest for experimentation. Father Busa's *Index Thomasticus* is frequently cited as the very first example of this line of research. The index emerged from Busa's PhD research, which focused on the concept of presence in the works of Thomas Aquinas. Busa's monumental efforts resulted in a set of tools which can be used to perform quantitative linguistic analyses of the complete oeuvre of Thomas Aquinas, which spans over one hundred titles.[69] Scholars after Busa have also used computational techniques to create concordances, to determine the likely authorship of unattributed works, or to characterise the stylistic features of collections of texts, among other purposes. Susan Hockey has argued that the use of electronic texts can effect "a real transformation in the way that scholars go about their work as new tools are introduced and new questions asked".[70] Studies are traditionally

[67] Walter Ong, *Orality and Literacy: The Technologizing of the Word*, p. 129.
[68] Ivar Armstrong Richards, *Principles of Literary Criticism.*, p. 1.
[69] Roberto Busa, "The Annals of Humanities Computing: The Index Thomisticus", in: *Computers and the Humanities*, 14 (1980).
[70] Susan Hockey, *Electronic Texts in the Humanities: Principles and Practice* (Oxford; New York: Oxford University Press 2000), p. 171.

limited to what is practicable to do by hand, but when scholars manage to capture the recognition of specific features of interest in algorithms, they can generally move beyond the established canon of literary works, and extend both the scale and the context of their research questions. Katherine Hayles notes that "the single most important issue in effecting transformation is scale".[71] In their *Digital Humanities Manifesto*, Pressner and Schnapp similarly observe that the field of digital humanities operates with an "economy [which] is abundance based". The authors coined the term "big humanities" to refer to forms of humanities research which exploit the "overflowing bounty of the information age" and which construct the "bigger pictures out of the tesserae of expert knowledge".[72]

1.5. Literary informatics and algorithmic criticism

Recent debates about the confluence of computing and literary studies have been dominated profoundly by the writings of Franco Moretti. In his essays "The Slaughterhouse of Literature" and "Conjectures on World Literature", which were first published in 2000, Moretti emphasises that the traditional scholarly method of close reading is inadequate for the examination of genres or literary periods in their entirety, as the sheer quantity of the texts that must be read exceeds what individual human readers can accomplish within a lifetime. As a result, conventional research focuses in on "a canonical fraction",[73] and establishes a remnant of ignored titles which Margaret Cohen refers to as "the great unread".[74] Moretti originally envisaged distant reading as a collaborative form of research, in which data collections produced by scholars dispersed over different locations and different disciplines are amassed and synthesised. By stitching a "patchwork of other people's research", studies in the field of literary history can eventually extend their scope and their ambitions. Distant reading thus entails a derivative line of research, which takes place "without a single direct textual reading". [75] According to Moretti, such a dissolution from the text itself is needed to ensure that the research can focus on diachronic or synchronic developments in the popularity of specific literary devices or genres.

During the decade that followed its initial articulation, the concept of distant reading proved highly influential. While the term was coined initially to stress the importance of data reuse, it was recognised increasingly that the distance that was proposed could likewise be achieved via the algorithmic manipulation of literary

[71] Katherine Hayles, *How We Think: Digital Media and Contemporary Technogenesis* (Chicago: The University of Chicago Press 2012), p. 27.
[72] Jeffrey Schnapp, Peter Lunenfeld & Todd Pressner, *The Digital Humanities Manifesto 2.0*, (2009), p. 4.
[73] Franco Moretti, *Distant Reading* (London: Verso 2013), p. 47.
[74] Margaret Cohen, "Narratology in the Archive of Literature", in: *Representations*, 108:1 (2009), p. 59.
[75] Franco Moretti, *Distant Reading*, p. 48.

works. Scholars who adopted the term found that it could felicitously be used as a blanket term for many of the existing methodologies of digital humanities research.[76] At present, the term is used most commonly to refer to computer-based methods which extract quantitative data from text corpora and which represent the results of data analyses in an abstract, non-textual manner. Since the term "distant reading" was originally coined, in a polemical fashion, as an alternative to the New Critical method of close reading, it is often assumed that the term implies an a priori rejection of the central objectives of close reading and of its concomitant attention to detail. This text will assume, however, that studies which adopt the method of distant reading do not necessarily disavow a detailed examination of individual texts, and that the methods of distant reading and close reading may also be used in conjunction.

The numerous panegyric depictions of computer-based research have, unsurprisingly, been the object of equally fierce criticism. A common critique is that, whereas literary research is centrally concerned with the interpretation of ambiguous and multi-layered texts, digital instruments principally support quantitative analyses of the more trivial aspects of literary works. Katie Trumpener considers statistical analysis to be "a relatively blunt hermeneutic instrument"[77] and argues that the human literary scholar will inevitably be needed to interpret the results of algorithmic processing. Stephen Marche similarly views the attempt to transform literature into discrete data as an act of sacrilege, and argues that the complicated meaning that generally inheres in works of artistic creation cannot be reduced to one-dimensional data. He also argues that questions of literary criticism cannot meaningfully be answered through the statistical analyses of data about texts, as such approaches invariably demand disproportionally narrow definitions of terms and unwarranted simplifications.[78] It can also be observed that, despite several decades of extensive experimentation, ICT tools have not yet managed to become part of the standard toolset in mainstream literary studies. Mueller notes that a small number of scholars "use computational methods extensively, but their work has had virtually no impact on major disciplinary trends".[79] Willet concurs that it is "difficult to find scholarly articles that cite electronic text collections as sources, or discuss the methodology of creating or using e-texts, outside of journals

[76] Matt Erlin & Lynne Tatlock, "Introduction: "Distant Reading" and the Historiography of Nineteenth-Century German Literature", in: Matt Erlin & Lynne Tatlock (eds.), *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, 2014, p. 3.

[77] Katie Trumpener, "Critical Response: I. Paratext and Genre System: A Response to Franco Moretti", in: *Critical Inquiry*, 36:1, p. 170.

[78] Stephen Marche, "Literature Is Not Data: Against Digital Humanities", *The Los Angeles Review of Books*, 28 October 2012 <http://lareviewofbooks.org/essay/literature-is-not-data-against-digital-humanities> (22 June 2014)., n.pag.

[79] Martin Mueller, "Digital Shakespeare, or towards a Literary Informatics", p. 289.

for computing humanists".[80] It is fair to say that computer-based analysis, as a methodology with the academic field of literary studies, is currently still in its infancy.

Symptomatic for the incipient nature of computer-based analysis is the fact that there is currently no widespread consensus regarding the subfield's nomenclature. The *Blackwell Companion to Digital Literary Studies*, which was first published in 2007, suggests that the manifold types of projects which are described in the volume could be grouped under the heading that was used in its title. Mueller proposes "literary informatics", as a name for the approach, drawing an analogy with the field of bioinformatics. In an insightful article on computer-assisted studies of literature, dating from 1983, John B. Smith uses the term "computer criticism" to identify "a mode of criticism that arises from using the computer". He also confesses to an uneasiness with the term, as it suggests that "it is the computer that does the criticism",[81] questioning the agency of the human critic who employs the computer. Studies that use lists of common words to investigate the authorship or the stylistic features of texts are often subsumed under the general rubric "textual analysis", but Stephen Ramsay notes that authorship attributions studies or stylometric studies are epistemologically distinct from inquiries in which the results of algorithms are used in service of the interpretation of the text, which Ramsay rightly views as "the core activity of literary studies". For this reason, Ramsay makes a distinction between "computational textual analysis" and an approach which he refers to as "algorithmic criticism", which is "criticism derived from algorithmic manipulation of text".[82]

The essence of the approach that is demonstrated in this thesis lies in the fact that computational techniques are used to analyse literary texts. In a sense, any term which combines these two central concepts may qualify as a suitable appellation. While phrases such as "e-Criticism" or "computational literary research" may perhaps be suggested as monikers for the field, the phrase *literary informatics*, which was proposed by Mueller, has the advantage that it clearly fits in with existing practices in a number of other disciplines. In addition, it appears sufficiently inclusive for the entire field of literary research. Next to literary criticism, the term "literary studies" encapsulates distinct activities such as literary history, authorship attribution research and studies on literary theory. For all of these related activities, the computer can potentially be of assistance and of

[80] Perry Willett, "Electronic Texts: Audiences and Purposes", in: Susan Schreibman, Ray Siemens, & John Unsworth (eds.), *Blackwell Companion to Digital Humanities*, Oxford: Blackwell, p. 250.

[81] John B. Smith, "Computer Criticism", in: Rosanne G. Potter (ed.), *Literary Computing and Literary Criticism: Theoretical and Practical Essays on Theme and Rhetoric*, University of Pennsylvania Press 1989, p. 14.

[82] Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism* (Urbana: University of Illinois Press 2011), pp. 2–3.

relevance. This thesis will mostly use the term *literary informatics* to refer to the wide range of transformations that are introduced to the field of literary studies as a results of a growing relevance of digital information. Mueller explains that the term also has the advantage that it stresses a degree of continuity with activities that existed before the advent of computing. Informatics "depends critically on the capability of digital technology to store, manipulate and move large quantities of information with great speed and accuracy", [83] and, with this qualification, the term *informatics* can also be connected to activities such as the creation of analogue concordances, or to practices of monks who transformed the vulgate bible into its constituent parts so that it could be studied more systematically.

Literary informatics is an emerging field, which, in essence, seeks to explore if technologies for the analysis and the management of digital texts can be adopted usefully and meaningfully within the context of literary studies. This broad definition is understood to include the many attempts of humanities scholars to describe and to enhance the structure of natural language texts through various forms of encoding. Examples of such mark up techniques include COCOA and, most notably, the Text Encoding Initiative.[84] Many of the technologies which are used within literary informatics research have emerged originally within other disciplines. Studies which focus centrally on computational analyses of large corpora of plain texts often appropriate many concepts and methods from text mining, an area of research which develops and exploits "a collection of methods used to find patterns and create intelligence from unstructured text data".[85] Text mining is related, in turn, to data mining, as both fields aim to generate useful information from collections of data by identifying patterns and regularities. In the case of text mining, however, such patterns are found "not among formalised database records, but in the unstructured textual data".[86] Since texts in natural language are mostly rife with idiosyncrasies and ambiguities, text mining operations often commence with a conversion of the original source into an explicitly structured format, thus allowing for a more systematic analysis of the reorganised units. Text mining, in short, views texts predominantly as databases with a number of additional challenges.

Text mining is a broad term, which may be used to refer to all computer-based manipulations and analyses of texts in natural languages.[87] The term, in turn,

---

[83] Martin Mueller, "Digital Shakespeare, or towards a Literary Informatics", pp. 284–285.

[84] The Text Encoding Initiative is explained in more detail in the Glossary in Appendix 1, together with a large number of other technical and statistical terms.

[85] Louise Francis, "Taming Text: An Introduction to Text Mining", in: *Casualty Actuarial Society Forum*, (2006), p. 52.

[86] Ronen Feldman, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data* (Cambridge: Cambridge University Press 2007), p. 1.

[87] Sholom Weiss et al., *Text Mining Predictive Methods for Analyzing Unstructured Information* (New York: Springer 2004).

unites a broad range of more specialised subdisciplines.[88] One important area of research is *Natural Language Processing* (NLP).[89] Like text mining as a whole, NLP aims to extract data from textual materials, but a distinctive feature is that analyses are predominantly based on a deeper knowledge of the linguistic structure of texts. Earlier research in NLP has led to an improved understanding of the way in which the various linguistic aspects of texts can be described and identified automatically or semi-automatically. In recent decades, sophisticated tools have been developed for, amongst other purposes, the recognition of grammatical and syntactic categories, or for the conversion of inflected word forms into their dictionary forms. Current research largely seeks to build on these existing tools and focuses more particularly on the development of more advanced software for tasks such as named entity recognition, machine translation, or unsupervised summarisation. Crucially, these tasks all demand an understanding not only of the grammar and the syntax, but also of the logical structure and the semantic contents of the text.[90]

The field of literary informatics adopts ideas and digital techniques from other disciplines, but this process of adoption should not be uncritical. Concepts and techniques which originated within mathematics or within the natural sciences are often based on particular methodological assumptions, and these assumptions are not necessary compatible with humanistic objectives. The process of appropriation ought to be accompanied, for this reason, by an appraisal of the relevance or of the applicability of these tools. New technologies invariably need to be naturalised into their new scholarly setting. In some cases, such processes of integration also demand a renaming. At present, technologies that were developed within the field of text mining have been adopted broadly among digital humanists. Many scholars prefer to describe their methodology, nevertheless, using Moretti's coinage as "distant reading". The very fact that the activity is explicitly labelled as a form of reading, rather than merely a form of processing or a form of mining, suggests that there are certain commonalities with traditional scholarly processes. Katherine Hayles' monograph *How we Think* contains a similar proposal for an alternative name for the field of text mining. Like Moretti, Hayles stresses that computational analyses of data ought to be viewed as a form of reading. Hayles argues that the position that only human beings can read betrays a species-centric bias. She uses the term "machine reading" to refer to a form of reading in which computer

[88] Ian H. Witten, "Text Mining", in: Munindar P . Singh (ed.), *The Practical Handbook of Internet Computing*, Boca Raton; London: Chapman and Hall/CRC 1999.

[89] Opinions vary on the precise relationships between Text Mining and NLP. This thesis follows the characterisation offered by Weiss et al. and by Witten, who regard NLP as an area within Text Mining. Jurafsky, however, views NLP as an independent area of research. See Daniel Jurafsky & James H. Martin, *Speech and Language Processing* (Englewood Cliffs: Prentice Hall 2008).

[90] Sholom Weiss et al., *Text Mining Predictive Methods for Analyzing Unstructured Information*, pp. 3–4, and passim.

algorithms are used "to analyze patterns in large textual corpora where size makes reading of the entirety impossible".[91]

As noted, the introduction of new concepts is ideally preceded by a critical assessment of the appositeness of these novel approaches. Moretti's writings were provocative because they sanction the methods borrowed from science in a seemingly uncritical fashion, and because they contain a denigration of some of the central tenets of conventional literary criticism. One of the central assumptions in Moretti's *Graphs, Maps and Trees* is that the trends that can be observed in the historical development of literary phenomena can be elucidated by imposing abstract models developed within the natural sciences. The chapter on "Maps", for instance, uses the classification trees which are applied more commonly within evolution theory to account for the emergence and the disappearance of particular types of detective novels. Moretti's studies frequently aim to forge singular conclusive answers to their questions, through "the pursuit of a sound materialistic method, and of testable knowledge".[92] Analogous to Hempel's observation that scientific research centrally explains observations via a reference to a universal law, many of Moretti's essays aim to identify the irrefragable principles that underpin particular developments in literary history.

Such a eulogistic and uncritical appropriation of scientific principles is repudiated fiercely, however, by Stephen Ramsay. Most articulately in his monograph *Reading Machines*, Ramsay stresses that the field of literary criticism "operates within a hermeneutical framework in which the specifically scientific meaning of fact, metric, verification, and evidence simply do not apply".[93] Literary criticism forcibly reflects the "significant traits" of humanistic research that were identified by Costis Dallas. Amongst other characteristics, Dallas explains that humanities research is often hermeneutic, "narrative, textual and rhetorical; [...] judgmental [...]; and idiographic".[94] Because of these features, research cannot easily be "reduced to formal syllogisms (laws, explanations), as prescribed in positivism".[95] Questions of literary interpretation can seldom be addressed via a single indisputable answer. In many cases, a particular reading is valuable precisely when it advances a scholarly discussion about a text and when it manages to expose additional layers of complexity. Ramsay emphasises that literary critics crucially concentrate on interpretation, and, as was noted earlier, he argues that there is a need for an algorithmic criticism, in which digital methods can veritably "assist the critic in the unfolding of interpretative possibilities".[96]

---

[91] Katherine Hayles, *How We Think: Digital Media and Contemporary Technogenesis*, p. 70.
[92] Franco Moretti, *Distant Reading*, p. 155.
[93] Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism*, p. 7.
[94] Costis Dallas, "Humanistic Research, Information Resources and Electronic Communication", p. 210.
[95] Ibid., p. 211.
[96] Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism*, p. 10.

Tom Eyers notes that Ramsay's argumentation is based on the thesis that the constraints associated with computation are "paradoxically enabling".[97] The range of textual aspects that digital methods can operate on is blatantly limited, but, ironically, such a neglect of the characteristics which human readers would view as obvious can also lead to unexpected critical results. Ramsay's algorithmic criticism aims to capitalise on the chasm between computation and critical interpretation. According to Eyers, however, Ramsay does not explain explicitly how this gap between quantitative and qualitative analysis can be bridged. More importantly, it is still unclear whether or not the textual transformations that can be effectuated by the machine can genuinely lead to an amplification of critical possibilities. Admittedly, the case study which is discussed in the first chapter in *Reading Machines* does not compellingly illustrate the actual potential of algorithmic criticism. Ramsay discusses a study in which digital methods were used to identify the distinctive vocabulary of the six main characters in Virginia Woolf's novel *The Waves*. The analysis exposed a pattern in which the speakers could roughly be divided along gender lines. As the outcomes of this investigation of *The Waves* do not differ widely from the findings of earlier critical readings of the novel, the study primarily demonstrates that quantitative methods can be used to corroborate results obtained via close reading, which is often a more cursory and more impressionistic from of analysis. While this seems a valid use of computation, Ramsay's writing strongly suggests that algorithmic criticism may also inaugurate more momentous and more venturesome critical possibilities. Computational tools for the analysis and the visualisation of texts may be viewed as instruments. As microscopes and telescopes have broadened the bandwidth of human perception, and as they have plainly enabled researchers in the natural sciences to address new kinds of questions, it may be expected that computer-assisted forms of reading can likewise enable literary scholars to perceive qualities of the text which have been disregarded in studies which lack such instruments. One of the crucial challenges of algorithmic criticism is to accept methods which are rooted firmly in quantification and in formalism, and to develop these into a set of heuristic procedures which may convincingly install new hermeneutic approaches, by exposing unanticipated singular qualities of literary works.

The relatively limited impact of computer-based methods on mainstream humanities can be explained by a number of factors. The most important of these, perhaps, is the difficulty that the use and the development of digital humanities tools require a proficiency in two seemingly distinct fields. Researchers firstly need to appreciate the intricacies and the subtleties that are involved in studying cultural artefacts. Secondly, to apply digital resources and tools well, it is also necessary to grasp processes of computational reasoning and to master the logic of digital

[97] Tom Eyers, "The Perils of the "Digital Humanities": New Positivisms and the Fate of Literary Theory", in: *Postmodern Culture*, 23:2 (2013), n.pag.

methods. Successful solutions ultimately manage to merge an understanding of the aims and the methods of the humanities fields with the affordances of information and communication technology. The acquisition of technical skills demands an investment in time, and scholars are often hesitant to spend their valuable time and energy because of a concern about the return on this investment. According to Ramsay, the use of digital tools still remains at the periphery of literary research because there are too few powerful statements of the benefits of such research. Ultimately, the scepticism that digital humanities still elicits can only be overcome if the field can actually manage to produce inspiring illustrations of the ways in which computational analyses can foster interpretation.

1.6. Research question

Algorithmic criticism endeavours to bridge a gap between computation and criticism, and to reconcile the quantitative and realist orientation of the toolset[98] with the evaluative and interpretative approach of the field in which these methods are adopted. Using digital methods, texts can be analysed and visualised in a variety of ways. Following Ramsay, it can be assumed that the significance of algorithmic processing results, more concretely, from the fact that the innovative perspectives that can be produced may ultimately stimulate interpretation. Many aspects of such computational transformations are still poorly understood, however. There is a degree of incertitude, first, concerning the precise nature of the research data. Additionally, while quantitative data can be analysed using numerous statistical methods, little information exists about the ways in which such procedures can genuinely stimulate a hermeneutic engagement.

This thesis aims to make a contribution to the further development of the field of algorithmic criticism, which, as Ramsay writes, currently consists "only in nascent form".[99] This study is interested, moreover, in examining the ways in which computational methods may expand or restrict the more traditional critical methods in literary studies. Various authors have posited a dichotomy between digital methods on the one hand and the conventional close reading method on the other, and have argued that the former may serve as a corrective to some of the perceived shortcoming of the latter. Moretti antagonistically proclaims that close reading, as a "theological exercise" and as a "very solemn treatment of very few texts taken very seriously",[100] is wholly inadequate within the context of literary history. Matthew Jockers writes, in a similar vein, that "the sheer quantity of available data makes the traditional practice of close reading untenable as an

[98] Eyers writes that "we may go as far as call it a quantitative ontology". See Tom Eyers, "The Perils of the "Digital Humanities": New Positivisms and the Fate of Literary Theory", n.pag.
[99] Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism*, p. 81.
[100] Franco Moretti, *Distant Reading*, p. 48.

exhaustive or definitive method of evidence gathering".[101] This thesis aims to understand the possibilities and the limitations of algorithmic criticism by comparing this approach to the traditional close reading method. More specifically, this study concentrates on the following question: *How does the methodology and the epistemology of algorithmic criticism relate to that of literary research which is based on conventional close reading?* Answers to this central question will help to define the nature of the difference between analogue and digital textuality at large, and the differential limitations of both in particular.

The central research question of this thesis is based on the assumption that there are differences between traditional humanistic research on the one hand and computer-based research on the other. The adjective "traditional" is used in this thesis to refer to forms of research which are based mostly on analogue resources and whose results do not depend on the application of digital research tools. This thesis aims to avoid the implication, nonetheless, that there is necessarily a deep gap between scholars who have integrated computational methods within their overall scholarly methodology and scholars who have not. Whereas many authors have highlighted the fissure between humanistic scholars who do and do not make use of digital methods, such a polarisation is ultimately unproductive. Digital humanists sometimes claim that the work of scholars who dodge computation lacks a scientific precision, and that it is additionally based on unjustifiably small data sets. Digital humanities research, conversely, has frequently been attacked for its alleged disregard of critical theory and for its reductionist conceptualisations of humanistic questions.[102] It seems more beneficial, however, to embed digital humanities research more closely within a humanistic critical tradition, and to assume that new methods should largely serve the same objectives as conventional forms of research. Paul Jay notes that the humanities at large are in a perpetual state of flux and stresses that the very notion of "a traditional core humanities practice"[103] is misleading. The humanities may be viewed as a loosely connected group of disciplines which collectively aim to understand man's relation to the world and the ways in which these experiences have been captured in cultural artefacts.[104] According to Wilhelm Dilthey, the "Human Sciences" study "our total lived experience of the human world and its incommensurability with all sensory

[101] Matthew Jockers, *Macroanalysis : Digital Methods and Literary History* (Urbana: University of Illinois Press 2013), p. 9.

[102] See for example Jean Bauer, "Who You Calling Untheoretical?", in: *Journal of Digital Humanities*, 1:1 (2011).

[103] Paul Jay, *The Humanities "Crisis" and the Future of Literary Studies* (Palgrave Macmillan 2014), p. 143.

[104] The University of Stanford defines the humanities as "the study of how people process and document the human experience. Since humans have been able, we have used philosophy, literature, religion, art, music, history and language to understand and record our world". See <http://shc.stanford.edu/what-are-the-humanities> (21 August 2014)

experiences of nature".[105] The methodologies for studying cultural artefacts are evolving continuously, and the digital humanities are best viewed as an emerging additional method which humanistic scholars can use to accumulate new ideas about the primary sources that are investigated. Bearing this nuance in mind, it seems reasonable to assume, nonetheless, that there are differences between computer-based criticism and criticism based on conventional methods, and this thesis aims to characterise both the disparity and the continuities.

Algorithmic criticism may ultimately be of relevance to the full breadth of literary studies, which is a highly diverse academic field, encompassing many different theoretical or methodological approaches. It may be argued, nevertheless, that the core objective of literary research is to perform literary criticism. In this thesis, the latter term is used to refer to the broad range of scholarly activities which centrally aim to illuminate the meaning of literary works or which aim to evaluate their quality or their importance. Over the course of several centuries, a broad variety of opinions have been developed concerning the purpose and the nature of textual interpretation. According to Mario Klarer, literary theories generally aim to clarify the various methods which can be followed in literary criticism, concentrating crucially on its "philosophical and methodological premises".[106] Klarer argues that there are essentially four non-exclusive theoretical approaches in literary theory. Text-oriented approaches analyses works of literature exclusively by considering "internal textual properties",[107] largely neglecting external aspects such as historical or biographical factors. Text-oriented literary research additionally encapsulates the various assiduous activities which are generally needed to secure the authority of the text or to reliably ascertain the authorship of a literary text. Author-oriented approaches seek to explain properties of texts conversely though references to biographical factors or to stated intentions in materials such as diaries or correspondence. In reader-oriented approaches, the focus is mostly on the reception of literary works, and on the other ways in which literary texts may affect readers socially or psychologically. Context-oriented approaches generally study the relationships between literary works and the broader cultural and social environment in which these are produced.[108] One of the most influential context-oriented movements is literary history, which fundamentally endeavours to assign literary works to distinct literary periods, frequently on the basis of methods borrowed from historical research.[109] The manifold theoretical lenses can be applied cogently to study the work of individual authors, but they can be adopted equally in studies which follow a comparative approach, and which aim to juxtapose works by different authors, works from different literary genres (dra-

[105] Wilhelm Dilthey, *Introduction to the Human Sciences* (Princeton University Press 1989), p. 61.
[106] Mario Klarer, *An Introduction to Literary Studies* (London: Routledge 1999), p. 75.
[107] Ibid., p. 76.
[108] Ibid.
[109] Ibid., p. 91.

ma, prose or poetry) or works from different literary periods. Synchronic and diachronic comparative analyses can focus on works written in in the same language, but they may also consist of cross-cultural or cross-linguistic comparisons. Scholars engaged in literary criticism often need to carefully select the single approach or set of approaches which seems most adequate for the texts that are studied, and which seem most likely to yield valid results.

Given the expansive diversity in the approaches and the perspectives which may be chosen within literary research, it does not seem feasible to exhaustively study the ramifications of digital methods for the full breath of the field of literary studies. It was decided, for this reason, to confine the research in a number of ways. A first restraint is that the focus is primarily on poetry. One of the central objectives of this study is to understand the computer's capacity to support interpretation and evaluation, and, in the light of this goal, the choice to concentrate on verse seems justifiable. Poetry is distinctly a genre in which interpretation is often strenuous because of the deliberate ambiguities that can arise from multiplicities in meaning. Additionally, since many existing text analysis tools function most productively in the case of relatively long texts in which the referents of words are also relatively stable, it can be assumed that computer-supported analyses of short poetic texts critically involves a number of challenges. Next to posing difficulties, however, poetic texts generally have a number of qualities which seem amenable to algorithmic analysis. Aspects such as rhyme, metre, alliteration and assonance are often tractable computationally and can consequently be analysed statistically. Surprisingly, within the field of literary informatics, aspects of meter and prosody have not been studied extensively.[110] As a second restraint, this thesis focuses exclusively on poetry written in the English language. The theoretical framework of this thesis is consequently based largely on literary theories and concepts that have been developed by theorists working within the Anglo-Saxon critical tradition. Whereas these limitations with respect to genre and language can diminish the broader applicability of this study's central findings, this confinement is also necessary to allow for a sharper analysis. The results of this study ultimately need to be supplemented by those of similar studies concentrating on prose texts or drama texts, or on texts written in other languages.

In the following chapters, the phrase *literary informatics* is used to refer to all types of literary research which make use of computational methods, including literary history and stylometric research. The term *algorithmic criticism* is viewed as a hyponym of literary informatics, denoting computer-based literary research which aims to provide support for the interpretation of literary works. It describes

[110] David Hoover notes that "[m]etrical analysis, because of the inherent reliance of meter on pattern, is a natural area for quantitative study, though there has been less research in this area than one might have expected". See David Hoover, "Quantitative Analysis and Literary Studies", in: Susan Schreibman & Ray Siemens (eds.), *A Companion to Digital Literary Studies*, Oxford: Blackwell 2008, p. 523.

the manner in which algorithmic manipulations of texts may stimulate acts of criticism. This thesis treats both *machine reading* and *distant reading* as synonymous with text mining. Machine reading forms a central method within literary informatics, and is contrasted in this thesis with the close reading method.

1.7. Structure of this thesis

The central question of this thesis is answered over the course of nine chapters. Chapter 2 characterises literary research based on conventional close reading, discussing views associated with Practical Criticism, New Criticism and New Formalism. As this thesis focuses on the analysis of poetry in the English language, the chapter primarily describes critical approaches which have been followed within the Anglophone tradition. Chapter 3 gives an overview of the current state of the field of literary informatics. It examines the main functionalities of existing text analysis tools, and it describes a number of representative research projects which have analysed literary texts via such tools. Computer-based literary research is described at a more abstract level in Chapter 4, which concentrates on the nature of the research data that can be produced or collected by its practitioners. Using conceptualisations offered by theorists in the field of information science and e-Research, the chapter proposes a classification of the various types of research data, and it introduces terminology that can be used to describe some of their properties. The fifth chapter is a brief synthesis of the findings of the first four chapters, and establishes the main distinctive characteristics of literary research based on computational methods.

In this dissertation, the possibilities and the limitations of computer-based literary criticism have also been explored on a practical level. I have carried out a case study which concentrated, more specifically, on the capacity of the computer to stimulate the interpretation of English poetry. This study consisted of a quantitative critical analysis of the poetry of the Northern-Irish poet Louis MacNeice. One of the central aims of the case study was to contribute to an alignment of traditional practices and scholarship based on data processing, through the algorithmic quantification of literary devices which have often been disregarded in existing computer-based research. I have created software for the recognition of a number of widely used poetic techniques, such as rhyme, alliteration, onomatopoeia and allusion. I have also developed various methods for the visualisation of these devices. The results of the case study are discussed in Chapters 6, 7 and 8.

Chapter 9 examines the various ways in which visualisation technologies may be of relevance to the field of literary studies. Drawing from a number of theories about visualisation, the chapter investigates the nature and the function of graphic renditions of research data. In addition, it scrutinises the capacity of data visualisations to invigorate hermeneutic processes. On the basis of the research that was conducted for this thesis, a number of crucial differences and similarities

have been identified between conventional close reading and machine reading. The main changes and continuities are discussed in Chapter 10.