Affordances and Limitations of Algorithmic Criticism

Peter Verhaar

Affordances and Limitations of Algorithmic Criticism

PROEFSCHRIFT
ter verkrijging van
de graad van Doctor aan de Universiteit Leiden
op gezag van Rector Magnicus prof. mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
ter verdediging op dinsdag 27 september 2016
klokke 13:45

door

Peter Anton Frans Verhaar
geboren te Axel
in 1976

Promotores:

Prof. dr. A.H. van der Weel
Prof. dr. K.H. van Dalen-Oskam (Universiteit van Amsterdam)

Promotiecommissie:

Prof. dr. Y. van Dijk
Prof. dr. S. Schreibman (Maynooth University)
Prof. dr. D. Van Hulle (Universiteit Antwerpen)
Prof. dr. W.T. van Peursen (Vrije Universiteit Amsterdam)

Contents

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 Dirextions 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.

Chapter 2

Close Reading

2.1. History and definition

In his *ABC of Reading*, Ezra Pound famously characterised literature as “news that stays news”.[111] Literary texts, and poems in particular, typically evoke an intricate set of visceral responses within readers, producing a mysterious appeal which compels readers to revisit and to reinterpret them repeatedly. Computation can be viewed as a method that can be used to strengthen and to invigorate this process of news-gathering. Digital methods may enable scholars to study the multifarious qualities of works of literature in a highly systematic manner, and they may potentially expose textual properties that remain hidden when texts are studied via more conventional methods. This thesis aims to understand the novel mode of studying literature that is engendered by machine reading by comparing it to scholarship based on close reading, which may be viewed as the dominant method for analysing literary texts in the physical realm. As a first step, the current chapter describes the main qualities of the close reading method. These qualities will be contrasted with the possibilities produced by digital methods in the following chapters.

Close reading is a broad term, which is commonly used to refer to a deeply attentive type of engagement in which readers minutely scrutinise the vocabulary, the grammar and the literary techniques found within individual fragments. It can be used to refer to a particular mode of reading, as well as to a description of the results of this type of reading. On the basis of this capacious description of close reading, it may be surmised that it has already been practiced for several centuries in studies on ancient rhetoric, in biblical exegesis and in classical philology.[112] Andrew DuBois concurs that “reading and responding to what one reads is an ancient practice, of which there exists a library of examples ecclesiastical, ecstatic, dogmatic, incidental, and so on”.[113] Neil McCaw observes that the methods which are discussed in Aristotle’s *Poetics* and in the works of the Greek critic Longinus may equally be viewed as forms of close reading.[114] In this thesis, the term close

[111] Ezra Pound, *ABC of Reading* (London: Faber and Faber 1991), p. 29.

[112] David Schur, “An Introduction to Close Reading”, (1998).

[113] Andrew DuBois, “Introduction”, in: Frank Lentricchia & Andrew DuBois (eds.), *Close Reading: The Reader*, Durham N.C.: Duke University Press 2003, p. 1.

[114] Neil McCaw, *How to Read Texts: A Student Guide to Critical Approaches and Skills* (London: Continuum 2008), p. 15.

reading will be used primarily to denote the form of textual engagement which came to prominence during the twentieth century, and which has since had a decisive impact on the nature of literary research.

The origins of the modern conception of close reading are commonly traced to the method of practical criticism that was developed in the late 1920s by I.A. Richards. One of Richards' central tenets was that the critical assessment of a work ought to be based solely on "the words on the page",[115] and that literary interpretation ought not to be affected by any knowledge of the historical circumstances in which the text was produced or of the author's biography. In his monograph *Practical Criticism*, Richards discusses a method which emerged from a series of pedagogical experiments conducted at Cambridge in which he asked students to analyse texts without supplying any information on their authors, titles or dates of creation. Richards' aim in removing these paratextual aspects was to identify potential causes of misreading and to address "the chief difficulties of criticism". Such complications include an inability to apprehend the central meaning of the text, an inattentiveness to the sonic effects, the potential influence of "mnemonic irrelevances" such as personal memories and the penchant for producing "stock responses" when views and affections are already formed before the start of the reading process.[116] Richards opines that a slow, critical and unbiased form of reading was essential to ensure that readers can be fully susceptible to the nuances and the ambiguities that can be produced by literary techniques.

The method of practical criticism became deeply influential after its adoption by the New Critics. Jessica Pressman stresses, however, that New Criticism, like close reading, is a highly unclear term. It does not have a single manifesto, and there are no clear statements of the objectives of the movement.[117] The ideas that came to be associated with New Criticism were spawned by a loosely organised group of scholars and poets hailing from the Southern United States, including John Ransom, Cleanth Brooks and William Wimsatt. Whereas individual theorists have placed different emphases, the New Critics were largely united in their conviction that literary texts ought to be analysed as autonomous objects, and independently of their social, historical and political contexts. Works were treated mostly as "verbal artefacts that transcend their compositional occasions and context".[118] Wimsatt and Beardsley categorically reject critical approaches in which the author's stated intentions are used as a basis for an interpretation of a text. A literary text is "detached from the author at birth and goes about the world beyond

[115] Ivar Armstrong Richards, *Practical Criticism : A Study of Literary Judgment* (London: K. Paul Trench Trubner 1929), p. 4.

[116] Ibid., pp. 13–15.

[117] Jessica Pressman, *Digital Modernism: Making It New in New Media* (Oxford: Oxford University Press 2014), p. 12.

[118] Adam Piette, "Contempory Poetry and Close Reading", in: Peter Robinson (ed.), *The Oxford Handbook of Contemporary British and Irish Poetry*, Oxford: Oxford University Press 2013, p. 231.

his power to intend about it or control it".[119] New Criticism was a formalist type of criticism which concentrated predominantly on structural and formal textual aspects, such as the grammatical structure, diction and literary devices.[120] Many New Critical readings of literary works are based on the conviction that there ought to be an organic unity between the form and the central meaning of the text. Critics typically aimed to demonstrate that the various linguistic and literary signs of a text all work in unison to produce its total effect.[121]

DuBois explains that New Criticism consisted of a "radical response to arcane Indo-European philology" and to a "historical scholarship that seems more deeply interested in sociology and biography than in literature".[122] The literary research of the first two decades of the twentieth century concentrated for a large part on philology, literary history and "impressionistic belletristic commentary",[123] and texts were often viewed primarily as historical documents carrying information about historical developments. According to Alan Liu, the historicist approaches which were attacked by the New Critics were essentially based on a form of distant reading, culling "archives of documents to synthesize a "spirit" (Geist) of the times, nations, languages, and peoples".[124] In his influential essay "Criticism Inc.", John Ransom writes that literary research was in danger of becoming "a branch of the department of history", and maintains that critics "must be permitted to study literature, and not merely about literature".[125] The New Critics pressed for a form of literary criticism which concentrated mostly on the formal and rhetorical features of the text, rather than on the text's author or on the text's reception. Critics such as John Ransom and Cleanth Brooks in particular aimed to demonstrate, moreover, that texts can be investigated thoroughly and with intellectual rigour. Ransom envisaged an objective form of criticism which is "more scientific, or precise and systematic".[126] While the New Critics strongly opposed the cold rationalism of the sciences, viewing its objectives as antithetical to the nature of humanistic research, they generally aimed to gain legitimacy for their approach by propagating a form of textual engagement which is ostensibly as meticulous and as accurate as the procedures used within the natural sciences.[127]

[119] W.K. Wimsatt, *The Verbal Icon: Studies in the Meaning of Poetry* (Lexington: University Press of Kentucky 1954), p. 5.

[120] Jessica Pressman, *Digital Modernism: Making It New in New Media*, p. 17.

[121] Stephen Matterson, "The New Criticism", in: *Literary Theory and Criticism: An Oxford Guide*, Oxford: Oxford University Press 2006, p. 168.

[122] Andrew DuBois, "Introduction", p. 3.

[123] Miranda B. Hickman, "Introduction: Rereading the New Criticism", in: Miranda B. Hickman & John D. McIntyre (eds.), *Rereading the New Criticism*, Columbus: Ohio State University Press 2012, p. 10.

[124] Alan Liu, "Where Is Cultural Criticism in the Digital Humanities?", in: Matthew Gold (ed.), *Debates in the Digital Humanities*, University of Minnesota Press 2012, p. 492.

[125] John Crowe Ransom, "Criticism Inc.", in: *The Virginia Quarterly Review*, Autumn (1937), p. 589.

[126] Ibid., p. 587.

[127] Miranda B. Hickman, "Introduction: Rereading the New Criticism", p. 8.

After the close reading method had been consolidated across English departments and creative writing courses across the United States during the 1940s and 1950s, New Criticism increasingly lost its authority in the 1960s and 1970s. The decline of the New Criticism's dominance is often connected to the emergence of deconstructionist or post-structuralist theories, and to a growing dissatisfaction with the fact that the New Critics confined the literary canon to a small group of authors whose works can productively yield to ahistorical and formalist analyses. The predilection to concentrate on well-constructed and self-contained poetry led to the ennoblement of Modernist and metaphysical poetry, written predominantly by "white male" authors,[128] and to an indifference to literature produced by marginalised communities and ethnic minorities. Gallop notes that the New Critical anti-historical approach "has been persuasively linked to sexism, racism and elitism".[129] The fierce criticism of New Criticism precipitated a number of theoretical correctives. In response to the stalwart formalism of the New Critics, scholars such as Stephen Greenblatt and Frederic Jameson argued for the need to recognise the influence of historical circumstances, and their views materialised through the formation of New Historicism, which aimed to "combat empty formalism by pulling historical considerations to the centre stage of literary analysis".[130] Theorists associated with reader-response theory additionally critiqued the claim that the meaning of the text can be extracted exclusively by studying the text itself, and posited that meaning is a social construct, depending strongly on literary socialisation and on contingent ideas of what constitutes meaning.[131]

Despite the fact that New Criticism had become a superseded paradigm towards the end of the twentieth century, the close reading method, which the New Critics helped to develop and to disseminate, continued to be of scholarly relevance. While the New Critical dismissal of history and of politics have frequently been targeted critically, close reading in itself has rarely been opposed. Adam Piette emphasises that close reading remained a key activity within semiotic, deconstructionist and post-structuralist schools of criticism.[132] The intricate ambiguities and conflicts which are scrutinised in deconstruction, for instance, can only be disclosed after minute examinations of syntax, vocabulary, devices and

[128] Cecily Devereux, ""A Kind of Dual Attentiveness": Close Reading after the New Criticism", in: Miranda B Hickman & John D McIntyre (eds.), *Rereading the New Criticism*, Columbus: Ohio State University Press 2012, p. 218.

[129] Jane Gallop, "The Historicization of Literary Studies and the Fate of Close Reading", in: *Profession*, (2007), p. 181.

[130] Harold Veeser, *The New Historicism* (New York: Routledge 1989), p. xi.

[131] Clare Connors, *Literary Theory* (Oxford: Oneworld 2010), p. 49.

[132] Adam Piette, "Contempory Poetry and Close Reading", pp. 4–5.

structures.[133] Gallop, more strongly, refers to deconstruction as the "offspring" of New Criticism and claims that, rather than challenging the centrality of close reading, it "infused it with a new zeal".[134] Even in readings informed by critical theory, the need to "establish the intrinsic context of the literary object" remains pivotal, as, without a solid apprehension of the nature of the text, "all extrinsic moves (which are also contextual moves) are themselves suspicious".[135]

Importantly, the close reading method must not be equated automatically with the type of textual engagement which was endorsed by New Criticism, since, as was noted, the method had originally been established by British scholars associated with practical criticism. The New Critics "did an enormous disservice to close reading"[136] by denying the relevance of historical and biographical material. As is indicated by Piette, William Empson's critical analysis of Shakespeare's *Sonnet 73* in *Seven Types of Ambiguity* is enlivened appreciably by references to Shakespeare's personal life, to Puritan iconoclasm, and to ecclesiastical life during the English Reformation. While the New Critics do not explicitly explain why such use of historical materials is inadmissible, this inattention to the "historical imagination" crucially divested the close reading method of one of its "most vital source of energy".[137]

It must be stressed, nevertheless, that the New Critics were not fully anti-historical. In his preface to the 1968 edition of *The Well Wrought Urn*, Brooks concedes that poems "do not grow like cabbage, nor are they put together by computors [sic]". As a text is undeniably created by a human author, it can be relevant "to consider his ideas, his historical conditioning, his theories of composition, and the background, general and personal, which underlies his work". It is considered permissible, moreover, to base interpretations partly on "the response of the reader".[138] McCaw explains that the New Critics recommended a "layered approach",[139] in which an initial strong focus on the poem as an autonomous and independent construction can be followed by an explanation of the text, in which

[133] Deconstruction is a school of philosophy which is centrally concerned with the manner in which texts produce their meaning. Jacques Derrida stresses that words only produce meaning via their contrasts with other words. Although the various theorists associated with deconstruction, on some points, have differing views on its more concrete applications within literary criticism, deconstructionist critical readings typically aim to pursue the alternative ways in which a text can generate meaning, next to the dominant sense which is seemingly intended. Analyses, for this reason, often entail a detailed and a recurrent consideration of the text's oppositions, contradictions and omissions. See Alex Thompson, "Deconstruction", in: Patricia Waugh (ed.), *Literary Theory and Criticism*, Oxford: Oxford University Press 2006.

[134] Jane Gallop, "The Historicization of Literary Studies and the Fate of Close Reading", p. 182.

[135] Andrew DuBois, "Introduction", p. 8.

[136] Adam Piette, "Contempory Poetry and Close Reading", p. 231.

[137] Ibid., p. 233.

[138] Cleanth Brooks, *The Well Wrought Urn: Studies in the Structure of Poetry* (London: Dennis Dobson 1968), p. x.

[139] Neil McCaw, *How to Read Texts: A Student Guide to Critical Approaches and Skills*, p. 55.

data about the author or about the cultural context can be applied usefully. In interpretation of poems which are overtly political, such as Yeats's *Easter 1916* or Marvell's *An Horatian Ode upon Cromwell's Return from Ireland*, it seems virtually impossible to forego references to historical events. Form and language formed the centre of critical attention, nevertheless, and this focus was often at the expense of an interest in the social and cultural background of texts.

While New Historicism, in an important sense, managed to compensate for the New Critical lack of historical awareness, the approach has been accused, in turn, of neglecting the specificity of literary form,[140] and of focusing the theoretical lens too narrowly on texts as carriers of information. Jane Gallop claims that, while it is important to study texts within their historical and cultural contexts, the lack of attention to language and to form also blurs the distinction between literary criticism and historical research. Marjorie Perloff facetiously refers to cultural criticism as "social sciences without statistics".[141] Since the 1990s, a growing number of literary theorists have sought to reposition close reading as the focal point of literary criticism, while simultaneously drawing attention to the historical contingency of literary form. Terry Eagleton, for instance, advocates "a dual attentiveness", in which scholars are sensitive both to "the grain and texture of literary works" and to "cultural contexts".[142] The emerging New Formalist movement likewise fuses the objectives of New Criticism and New Historicism and recognises the simultaneous importance of close reading and of historical contextualisation. New Formalism aims to pay close attention to form "without succumbing to either the reactionary conservatism or the ahistorical and apolitical nature of New Criticism". At the same time, it aims to understand "the role form plays without compromising our understanding of history, cultural context, and the mandates of post-structuralist literary inquiries".[143]

Close reading is best viewed as a generic formalist method which can be employed equally by different schools of literary theory, albeit with varying implementations. Frank Lentriccia explains that, while the precise boundaries of close reading are uncertain, the "commitment to close attention to literary texture and what is embodied there"[144] forms a common ground for many theoretical orientations. Jane Gallop stresses that the essence of literary studies does not lie in the nature of the texts that are being read, but, rather, in the fact that it analyses texts via the method of close reading. Katherine Hayles stresses similarly that, after the New Critical hold on the literary canon was terminated, and after literary studies

[140] Miranda B. Hickman, "Introduction: Rereading the New Criticism", p. 3.

[141] Majorie Perloff, *Differentials: Poetry, Poetics, Pedagogy* (University of Alabama Press 2004), p. 13.

[142] Terry Eagleton, *How to Read a Poem* (Malden Mass.: Blackwell Pub. 2007), p. 8.

[143] Verena Theile, "New Formalism(s): A Prologue", in: Verena Theile & Linda Tredennick (eds.), *New Formalisms and Literary Theory*, Basingstoke: Palgrave Macmillan 2013, p. 12.

[144] Frank Lentricchia, "Preface", in: Frank Lentricchia & Andrew DuBois (eds.), *Close Reading: The Reader*, Durham: Duke University Press 2002, p. ix.

expanded its scope to include works of popular culture, close reading assumed "a preeminent role as the essence of the disciplinary identity".[145]

Despite the fact that there are marked differences between the forms of close reading that have been propagated by practical criticism, New Criticism and deconstructionist criticism, McCaw usefully argues that close reading can be defined by three central features. A first characteristic is that the method is primarily concerned with the text as an independent unit. Close reading, secondly, aims to illuminate the meaning of the text "through an examination of how it operates". A third central assumption is that the context of the text is of less importance than the language.[146] The first characteristic that is identified by McCaw — the notion that close reading takes place at the level of individual texts, or at the level of shorter fragments within individual texts — is particularly useful in distinguishing close reading from other modes of studying texts. Close reading mostly begins with the identification of occurrences of distinct literary devices or of noteworthy vocabulary, and its eventual objective is to analyse how these phenomena interact at the level of sentences, paragraphs or stanzas. At the level of these textual units, the various literary devices may reinforce each other, or they may cause striking conflicts or paradoxes. Formalist critical approaches such as structuralism and Russian Formalism, by contrast, were often interested in aggregations which exceeded the individual text. Smith explains that structuralist critics created abstractions of texts "with the aid of stratified levels of conceptual categories",[147] in order to investigate the linguistic characteristics of literary genres or periods in their entirety. The Russian Formalists likewise studied the linguistic aspects of works in order to contribute eventually to an understanding of the general laws and the *literariness* of literary language. Vladimir Propp, for instance, reduced formal aspects of individual literary works to instances of distinct categories in order to describe their generic principles. In one of his best-known studies, Propp classified the narratives contained in Russian fairy tales on the basis of 31 cardinal functions.[148] Formalist readings which aim to expose the broader patterns within

[145] Katherine Hayles, *How We Think: Digital Media and Contemporary Technogenesis*, pp. 57–58.

[146] Neil McCaw, *How to Read Texts: A Student Guide to Critical Approaches and Skills*, p. 56.

[147] John B. Smith, "Computer Criticism", p. 24.

[148] See Vladimir Propp, *Morphology of the Folktale* (Austin: University of Texas Press 1968). Similar conceptualisations were developed by Skaftymov, who viewed elements of the plot or features of literary characters as components within an overarching aesthetic structure and by Reformatsky, who concentrated on the structural relations between a work's themes, motives and plots. See John B. Smith, "Computer Criticism", p. 22.

large collections of texts partly foreshadow the aims of the approach which has been referred to more recently as "distant reading".[149]

Since the level of analysis forms an important distinctive characteristic of close reading, it is useful to introduce terminology that can be used to describe the two main levels that can be distinguished. In this thesis, the term "micro-level" is used to refer to the level of sentences, paragraphs or stanzas, at which literary scholars can observe individual textual units, such as words or literary devices, within their original context. Analyses at the macro-level, conversely, aim their attention at corpora consisting of multiple texts.[150] Potentially, a third plane of analysis may be distinguished in between the micro-level and the macro-level. Next to collecting data about large collections of literary works, scholars can also aggregate discrete data at the level of individual texts. Such operations can reveal aspects about the text as a whole, but they have the effect that scholars lose the ability to study textual units in their original context. This latter form of research will be viewed as a specific form of macro-analysis, however, as this thesis is mostly concerned with the differences between the focus on individual text fragments and the focus on abstract rendition of texts, created on the basis of quantitative data about such fragments.[151]

Following McCaw's concise conceptualisation, it may be claimed that close reading is centrally defined by two central activities. Close reading consists, on a first level, of a minute descriptive analysis of formal aspects such as syntax, vocabulary, diction and literary devices. It is based on a protracted attention to the form and to the language of the literary work. Jane Gallop stresses that close reading demands "looking at what is actually on the page, reading the text itself, rather than some idea 'behind the text'". The method demands the capacity "to

[149] Smith explains that this objective was achieved only partially, as many of the structural elements which are studied by structuralists were "never codified a set of methods or techniques that is adequate and general enough to accommodate close, sophisticated analyses of a variety of literary works" (p. 15). Many of the structuralist schools are defined by "the impracticality of applying their perspectives to large, full length texts" (p. 25). See John B. Smith, "Computer Criticism".

[150] The definition of the micro-level and the macro-level differ slightly from the way in which these terms have been defined by Matthew Jockers. According to Jockers, micro-analyses focus on aspects of a single text, meso-analyses concentrate on small text corpora, and macro-analyses explore properties of large text corpora. See Matthew Jockers, *Text Analysis with R for Students of Literature* (Springer, 2014), p. 4. As I assumed that it can be difficult to make a sharp and consistent distinction between small corpora and large corpora, the terms "micro-analysis" and "meso-analysis" have been redefined.

[151] It must be noted, also, that the definitions which have been given partly hinge on the definition of the term "text". If a text can be a short story, an examination of a collection of short stories would form an example of an analysis at the macro-level. Conversely, there may also be reasons for viewing the full collection of stories as a single text. This text will abstract from such complications, however. In cases where there may be confusion, the context will clarify the signification of these terms as much as possible.

read NOT what SHOULD BE on the page but what IS".[152] A second core activity can be referred to as interpretation. The aim of interpretation is generally to illuminate the meaning of a text, but, importantly, in the case of literary works, it also focuses on the manner in which the various formal features of a text contribute to the text's general meaning. Importantly, critics can be interested both in the confluence and in the conflicts between form and meaning. While New Critics have claimed that form and semantics need to cohere organically, deconstructionist critics are primarily attentive to the collisions that can arise between the language and the message that is conveyed by this language. These two central activities, descriptive analysis and interpretation, will be discussed in more detail in the following section. With respect to close reading, the act of evaluation may potentially be identified as a third activity. Close reading can help scholars to make a critical assessment of the literary quality of a text. Evaluation will not be viewed as a core component of close reading, however, but as an additional objective which the method of close reading is expected to support. This chapter closes with a brief section about the qualitative assessment of works of literature.

2.2. Components of close reading

2.2.1. Descriptive analysis

According to Roman Jakobsen, literary texts have a "poetic function" which refers to the "set (*Einstellung*) towards the message as such". There is frequently a "focus on the message for its own sake".[153] New Critics have often stressed that because of the importance of form, literary works cannot be paraphrased. Cleanth Brooks stresses that poetry must be considered as a structure, in which the various components have been arranged meticulously in order to produce a cumulative effect. While it is possible to describe what the poem is generally about, such a paraphrase is not "the real core of meaning which constitutes the essence of the poem".[154] Literary texts have a "meaning that cannot be made by other means".[155]

A close reading of a literary work often commences with an examination of the text's linguistic aspects and of the literary devices that have been used. Marjorie Perloff explains that literary research can be viewed as "a branch of rhetoric". Rhetoric concentrates on the manner in which a text is composed, and, within literary criticism, this mainly entails "the examination of diction and syntax,

[152] Jane Gallop, "The Ethics of Reading", in: *Journal of Curriculum Theorizing*, (2000), pp. 7–8. Capitals are in the original.

[153] Roman Jakobson, "Closing Statement: Linguistics and Poetics", in: Thomas A Sebeok (ed.), *Style in Language*, Advances in Semiotics, Cambridge, Mass: MIT Press 1960, p. 356.

[154] Cleanth Brooks, *The Well Wrought Urn: Studies in the Structure of Poetry*, pp. 158–160.

[155] David Schur, "An Introduction to Close Reading", n.pag.

rhythm and repetition, and the various figures of speech".[156] The identification of these core properties eventually forms the basis for more sagacious analyses. This section provides a brief synopsis of the textual phenomena which can be examined during descriptive analyses. The scope will be limited to characteristics of poetry, for two reasons. The New Critics were interested in texts which displayed instances of irony, paradox and ambiguity, and, because of this aim, many of the New Critics were predominantly concerned with poems, which typically "traffic in disruption and disorientation".[157] This section also places a special emphasis on the close reading of poetry because of the fact that the case study presented in this thesis centres around a corpus consisting of poems. As it is impossible to do full justice to the manifold ways in which scholars have investigated poetry, however, this overview does not aspire to be exhaustive. While the description of the literary phenomena that follows may additionally be perceived as reductive or as somewhat trite, the main aim of this section is to develop an elemental framework which can be used in subsequent chapters as a basis for a comparative analysis of traditional practices and computational approaches.

Costas Dallas notes that research projects in the humanities commonly start with an "[i]dentification of the activity or product to be explained, and resolution into elements". The elements which are identified are subsequently described "in terms of the 'language' of the discipline at hand".[158] The discipline of literary criticism has devised an elaborate system of terms which may be used to classify particular textual aspects, and, in agreement with Dallas' observations, analyses of poetry often consist of the isolation of particular textual phenomena for closer inspection, and of the subsequent application of literary terms. Piette explains that "close reading is a habit of attention to the ways the different kinds of material come together in the formal design" and that the analysis "simply separates out the elements so they become plainer to see".[159] The descriptive analysis of a literary work typically consists of the recognition of a textual element as an instance of a particular literary device. In this thesis, the term "literary device" will be used as "an all-purpose term used to describe any literary technique deliberately employed to achieve a specific effect".[160]

An extensive range of terms is available, for example, for describing the elements that can be identified during a prosodic analysis. Prosody, more specifically, is the study of sonic and rhythmic characteristics, and it entails the examination of rhyme, rhythm and metre. Phenomena such as end rhyme and metre crucially come into existence as a result of the fact that the poetic text is

[156] Majorie Perloff, *Differentials: Poetry, Poetics, Pedagogy*, p. 6.

[157] Jessica Pressman, *Digital Modernism: Making It New in New Media*, pp. 14–15.

[158] Costis Dallas, "Humanistic Research, Information Resources and Electronic Communication", p. 211.

[159] Adam Piette, "Contempory Poetry and Close Reading", p. 238.

[160] "Device", in Chris Baldick, *The Oxford Dictionary of Literary Terms* (Oxford: 2009), p. 85.

divided into separate lines.[161] Mary Oliver explains that the word "verse" derives from a Latin word signifying "to turn".[162] By turning the various verse lines, the poet establishes particular linguistic units, and can begin to craft phonetic and rhythmic patterns within lines and across lines. Verse lines which are written in accentual-syllabic metre generally have a regular number of stressed and unstressed syllables, and such lines can often be classified by considering the type of verse feet that are used (e.g. iamb, trochee, spondee, dactyl) and the total number of feet in each line (e.g. trimeter, tetrameter, hexameter).[163] The term "rhythm" is used to refer to the overall speed of the verse lines. Eagleton describes rhythm as one of the most "primordial" of poetic features. While metre supplies a regular pattern of stressed and unstressed syllables, rhythm often varies from line to line. The rhythm of a verse line can be determined by the use of pauses such as line endings or caesura, and by alterations of long vowels, short vowels and consonant clusters. If a line mainly consists of short vowels and single consonants, in mono-syllabic words, the rhythm is generally experienced as fast. Rhyme, thirdly, is a very familiar technical device in poetry. It consists of "a unity of identity and difference".[164] Lines which rhyme perfectly share final phoneme sequences. When there is only an agreement in the sounds of consonants or of vowels, such agreements are referred to as pararhymes or slant rhymes.

A broad range of terms are likewise available for the description of the form of a poem. While poems can be stichic, meaning that there is simply a sequence of verse lines,[165] many poems are divided into stanzas. Stanzas can be characterised by considering the number of lines, the rhyming schemes and the metrical patterns which are used within these stanzas. One example of a two-line form is the heroic couplet, which consists of two rhyming iambic pentameters. Three line-forms can either be triplets, in which all lines rhyme, or tercets, in which one or more lines do not rhyme. Four-line types may be single-rhymed, cross-rhymed, couple-rhymed, among other types. These basic forms can be combined into forms which contain larger number of lines, such as sonnets, villanelles, sestinas or octava rima. Poems can also have an open form, which means that the form is variable.[166]

[161] Eagleton defines a poem as "a fictional, verbally inventive moral statement in which it is the author, rather than the printer or word processor, who decides where the lines should end", see Terry Eagleton, *How to Read a Poem,* p. 25. This particular description, which may sound slightly pedestrian, implies that features such as rhyme or meter do not serve as defining characteristics of poetry. Eagleton explains that there are many poems which lack any rhyme or rhythms, while there are simultaneously many examples of prose texts in which poetic techniques such as rhyme or alliteration are used abundantly.

[162] Mary Oliver, *A Poetry Handbook* (San Diego: Harcourt Brace & Co. 1994), p. 35.

[163] The description or the visualisation of the metre is generally referred to as "scansion".

[164] Terry Eagleton, *How to Read a Poem*, p. 132.

[165] John Lennard, *The Poetry Handbook: A Guide to Reading Poetry for Pleasure and Practical Criticism* (Oxford [u.a.]: Oxford Univ. Press 2005), p. 24.

[166] Ibid., pp. 23–33.

In classical rhetoric, a distinction is often made between “tropes” or “figures of thought”, which are literary devices in which “words or phrases are used in a way that effects a conspicuous change in what we take to be their standard meaning”, and “figures of speech” which “depart from what is experienced by users as standard, or literal, language mainly by the arrangement of their words to achieve special effects”.[167] This distinction, together with the associated terminology, is often contested, however.[168] Baldick states that the term “figurative language” can be used to refer either to “[a]n expression that departs from the accepted literal sense or from the normal order of words”, or to one “in which an emphasis is produced by patterns of sound”. Devices such metaphor, metonymy, simile and personification may be viewed as examples of devices based on shifts in meaning. Devices such as assonance, consonance and alliteration are centrally based on repetitions of sounds. A large number of literary devices produce emphasis through the placement or the repetition of words or of sections of words, such as anaphora, chiasmus or polyptoton.

Analyses of poetry may also concentrate on their diction or on their syntax. Diction refers to the words which are chosen to express a particular message, including the reason for and the consequences of such choices. Diction can be classified as formal or colloquial, as concrete or abstract, or as complicated or simple. Words may be of a Germanic or of a Romance origin, and they may be polysyllabic or monosyllabic.[169] In poetry, the demands of metre and rhyme often place restrictions on the vocabulary. Words typically belong to a particular register of speech. The words in a text are often taken from the same register, but, when different registers are combined, this often draws attention to particular words.[170] The syntax of a text, furthermore, may be “clear or unclear”, or “verbose or economic”. Analyses may concentrate on occurrences of particular syntactic constructions, such as split infinitives, passive and active constructions,[171] or on the occurrences of personal pronouns. In stylistic research, it can be revealing to study shifts in perspective, such as that from a first person singular to a second person singular. In poetry, the syntax is often deliberately complicated. The meaning of a sentence may be confounded because of an unconventional word order, or because of the fact that the part of speech of individual words are unclear. Syntax, as such, can clearly contribute to the overall ambiguity of poetic texts.

[167] Meyer Howard Abrams, *A Glossary of Literary Terms* (Fort Worth: Harcourt Brace Jovanovich College Publishers 1993), p. 344.

[168] Chris Baldick, *The Oxford Dictionary of Literary Terms*.

[169] John Lennard, *The Poetry Handbook: A Guide to Reading Poetry for Pleasure and Practical Criticism*, pp. 103–105.

[170] *The Princeton Encyclopedia of Poetry and Poetics* (Princeton: Princeton University Press 2012), p. 358.

[171] John Lennard, *The Poetry Handbook: A Guide to Reading Poetry for Pleasure and Practical Criticism*, pp. 120–121.

In his monograph *How to Read a Poem*, Eagleton discusses a number of additional terms which may be used to characterise poetry. The mood of a text, first, describes its general atmosphere. The text's tone refers more specifically to the manner in which this atmosphere is expressed. It is the general attitude which is conveyed. According to Eagleton, a tone can be "exultant", "jubilant", "bombastic", "arch, abrupt, dandyish, lugubrious, rakish, obsequious, urbane, exhilarated, imperious".[172] "Volume" refers to the loudness or the softness of a line. The presence of many exclamation marks may indicate a high volume. The "intensity" of a poem refer to the density of particular devices. The intensity of a poem is frequently experienced as high when it contains many literary devices which are based on forms of repetition, such as alliteration, assonance, internal rhyme or polyptoton. The texture, finally, is the degree to which "a poem weaves its various sounds into palpable patterns".[173] Describing the texture demands attention to occurrences of sharp consonants such as plosives and softer sounds such as nasal consonants, fricatives and vowels. Eagleton notes that many of the aspects which characterise the style of an author are difficult to formalise.

2.2.2. Interpretation

Close reading often focuses intimately on the language of a literary work. A text invariably has a particular meaning, however, and, an obdurate focus on questions of form "downplays the cognitive import".[174] Next to analysing the form of the text, literary scholars also aim to illuminate the meaning of the text. An investigations of the form is usually regarded as being in the service of the overall illumination of the text's meaning. In a narrow sense, interpretation entails the identification of the theme of a work. A text often describes a specific atmosphere of specific events, but the words of a text typically epitomise more recondite or more abstract concepts at a higher level of abstraction. A theme may be defined as "a salient abstract idea that emerges from a literary work's treatment of its subject-matter".[175] Themes do not consist of paraphrases of the plot or of the images which are evoked. According to Robert Scholes, themes represent "a great cultural code" or a "great cultural axis". They are "the generalised oppositions that structure our cultural systems of values". They are mostly described using abstract terms such as "love", "war", or "decay". Robert Scholes argues that the themes of a literary work can often be found by considering the repetitions and oppositions which are evoked in a work. [176] Willy van Peer concurs that themes commonly reflect widespread cultural

[172] Terry Eagleton, *How to Read a Poem*, p. 116.

[173] Ibid., p. 120.

[174] Majorie Perloff, *Differentials: Poetry, Poetics, Pedagogy*, p. 7.

[175] Chris Baldick, *The Oxford Dictionary of Literary Terms*.

[176] Robert Scholes, *Textual Power: Literary Theory and the Teaching of English* (New Haven: Yale University Press 1985), pp. 133–134.

anxieties, connected to particular social changes.[177] He adds that they are generally described in a "foregrounded situation" and that they are emotionally charged. An important characteristic of themes, furthermore, is that they tend to resurface in different cultures and in different historical periods.

In *Understanding Poetry*, Brooks and Warren stress that an apprehension of the meaning of the text does not exclusively consist of a description of the theme of the text. Whereas the theme, being the central idea of the poem, can mostly by summarised in a single statement, the meaning of the poem is the "basic attitude and idea implied by a poem when it is understood as a whole". Through elements such as mood, tone, diction and imagery, the poet can express a particular emotional response to the theme. Through the rich poetic language, the author aims to convey the "special import of the dramatization of a situation". Brooks and Warren suggest that interpreters ought to be fully susceptible to the effects which are elicited by the interfusion of literary techniques. The meaning can be grasped by "witnessing and taking part in the great human effort to achieve meaning through experience".[178] Northrop Fry explains analogously that literary texts contain complicated semantic fields, which produce effects on many different levels. To fully appreciate the meaning of the text, literary critics need to engage in a highly immersive and attentive form of engagement, and need to be willing to surrender "the mind and senses to the impact of the work as a whole".[179]

The linchpin of the connection between form and content is the presumption that literary devices can have particular connotations and that they can produce particular effects. An iambic metre, for instance, is commonly experienced as exuberant and cheerful. Falling metrical feet, such as dactyls or trochees, may be said to have a negative or a melancholy connotation.[180] Eagleton notes that pararhymes can produce "mourning, haunting, almost eerie" effects. Literary forms may likewise be connected to specific expectations. Sonnets, for instance, are traditionally "love poems and declarations of courtship", the ottava rima is often thought of as comic, and tetra-metric couplets are conventionally regarded as "epic and serious". [181] An examination of the literary devices that were found during the descriptive analyses may also reveal that different types of literary devices produce effects which are very similar. The haunting effects that are produced by pararhymes, for instance, may be reinforced within a poem by its use of unconventional

[177] Willie van Peer, "Where Do Literary Themes Come From?", in: Max Louwerse & Willie van Peer (eds.), *Thematics: Interdisciplinary Studies*, Amsterdam/Philadelphia: John Benjamins 2002, pp. 255–260.

[178] Cleanth Brooks, *Understanding Poetry* (New York: Holt Rinehart and Winston 1960), p. 267.

[179] Northrop Frye, *Anatomy of Criticism* (Princeton, New Jersey: Princeton University Press 1957), p. 77.

[180] John Lennard, *The Poetry Handbook: A Guide to Reading Poetry for Pleasure and Practical Criticism*, p. 6.

[181] Ibid., p. 23.

syntax. As these effects of literary devices often depend on their usage within a particular context, the connotations or the effects of devices are difficult to formalise in logically consistent rules.

Interpretations can be constructed, subsequently, by connecting the patterns that emerge from an analysis of the effects of literary devices to the central themes of the text. David Schur surmises that a literary work consists of "underlying thoughts that have been converted into forms". The relation between form and contents is circular, moreover, as literary authors convert "thoughts into forms and forms into thoughts".[182] The overarching theme can help interpreters to read particular details, and the details of the text may inversely affect the understanding of the general purport of the work. One of the aims of the interpreter may be to demonstrate that the different strata of the text collectively develop a coherent set of ideas.

Brook's and Warren's suggestion that literary interpretation demands "sympathetic imagination"[183] on the part of the reader is strongly reminiscent of the hermeneutic philosophy of Hans-Georg Gadamer. Interpretation, according to Gadamer, is based on a pre-reflective or non-theoretical form of understanding which differs profoundly from the form of understanding that prevails within the natural sciences. The objective of interpretation is not to extract a singular objectively correct meaning, detached from the person who performs the interpretation. A hermeneutic engagement typically consists of a dialectical process, in which a reader, with unique interests and preconceptions, responds to the particularities and the singularities of the text. The result is a shared product, in which the reader's interests and predilections form an integral part of the meaning that is constructed. The manner in which the meaning ensues is not necessarily bound by an internal logic.[184] Gadamer makes an important distinction between knowing and understanding.[185] Knowing demands that there is a reliable point of view from which the text can be viewed in an objective perspective. The interpretation of a literary work, by contrast, demands an understanding, which arises when the text produces "an increased self-knowledge and insight" on the part of the reader. The main consideration is "whether the interpretation is itself productive or not, whether it opens up new dimension of thought and new lines of inquiry". The validity of the interpretation cannot be assessed separately from the interpreter. A reading may be considered valid if it leads to an "increased, or more productive self-understanding".[186]

[182] David Schur, "An Introduction to Close Reading", n.pag.

[183] Cleanth Brooks, *Understanding Poetry*, p. 267.

[184] Timothy Clark, "Interpretation: Hermeneutics", in: Patricia Waugh (ed.), *Literary Theory and Criticism: An Oxford Guide*, Oxford University Press 2006.

[185] David Hoy, *The Critical Circle: Literature, History, and Philosophical Hermeneutics* (Berkeley: University of California Press 1978), p. 46.

[186] Ibid., p. 49.

The close reading method, and, particularly its interpretative components, invariably demands subjective judgements. During the descriptive analysis, the decision to concentrate on specific elements and to disregard certain other elements is typically based on individual preferences. Texts can be read and interpreted in many different ways. Gadamer stresses, furthermore, that understanding is inescapably rooted within a particular historical situation. An interpretation arises out of a mediation between the text to be interpreted and the historical standpoint of the reader.[187] Different generations reads texts differently, and there "cannot, therefore, be any single interpretation that is correct 'in itself'".[188] A recognition of the historicity and the subjectivity of interpretations appears to lead to a relativism, in which it is impossible to compare the validity of different interpretations on rational grounds. Gadamer underscores, nevertheless, that the interpreter has the obligation to follow the text faithfully and to refrain from actively projecting idiosyncratic ideas onto the text.[189] The fact that the act of interpretation cannot be explained or formalised via an encompassing theory does not mean that it is irrational. Critics ought to describe the unique qualities of the text faithfully, and ought not to rebuild these according to personal insights.[190]

Eagleton argues in a similar vein that, whereas the aspects which are discussed in an interpretative reading rarely have an explicit presence in the texts, these are not completely arbitrary. A critic cannot make the words on the page "mean anything", as the words in a language have meanings which, to some extent, are codified. Word meanings, including both denotations and connotations, are constructed socially. Interpreting a text is "a rule-governed social practice". At the same time, readers are not "inexorably bound by these built-in interpretations".[191] While there is generally a large degree of latitude, words are often bound to a delineated cluster of associated meanings, and a reading can only be perceived as valid if it bases itself on these shared concepts of signification, rather than on deeply personal associations or on purely subjective preferences.

2.3. Evaluation

Next to an analysis of the language of the text and a consideration of the relation between the form and the meaning, critics may also determine whether or not a text has literary value. Robert Scholes refers to this latter activity as criticism proper. A descriptive analysis results in a "text within text", interpretation results in a "text upon text" and the aim of criticism is to produce "text against text".[192]

[187] David Hoy, *The Critical Circle: Literature, History, and Philosophical Hermeneutics*, p. 52.

[188] Hans-Georg Gadamer, *Truth and Method* (New York: Seabury Press 1975), p. 398.

[189] David Hoy, *The Critical Circle: Literature, History, and Philosophical Hermeneutics*, p. 67.

[190] Ibid.

[191] Terry Eagleton, *How to Read a Poem*, p. 109.

[192] Robert Scholes, *Textual Power: Literary Theory and the Teaching of English*, p. 24.

While description and interpretation basically result in a clarification of the work in itself, critics can additionally evaluate the literary quality of a work, and this often demands an extrinsic move, in which the qualities of the work are assessed on the basis of extraneous criteria.

The critical debates concerning the value of literary authors or of literary texts have often focused on the question whether or not it is possible to establish objective grounds for aesthetic judgements.[193] Many scholars associated with practical criticism and with New Criticism have claimed, implicitly or explicitly, that this is possible, and have striven to define the observable properties that determine literary quality. William Epson states that literary works can merit scholarly attention if they can yield to analyses which are intent on exploring multiple, often contradictory, meanings, and, Cleanth Brooks stresses, along similar lines, that poems are valuable if they make use of "the language of paradox" which juxtaposes ideas or connotations which seem incompatible. F.R. Leavis was "virtually obsessed with deciding what did and did not belong in the canon of 'great texts' worthy of further study".[194] In *The Great Tradition*, Leavis rather aggressively declares a list of the "novelists in English worth reading".[195]

A number of scholars have argued, to the contrary, that evaluative assessments can impossibly be motivated objectively. Terry Eagleton stresses that the criteria which are used to establish literary value are inevitably constructed within a particular social and cultural setting. The literary work is not "valuable in itself, regardless of what anyone might have said or come to say about it".[196] Northrop Frye even surmises that, since evaluation cannot be objective, it ought to be avoided by critics. He claims that literary criticism ought to base itself exclusively on observable properties and verifiable claims, and notes that, because there are "no facts" in "the history of taste, [...] the history of taste has no organic connection with criticism".[197] If it is accepted that evaluation is ultimately subjective in nature, the aim to establish the literary value of a work also seems in conflict with Wimsatt's and Beardsley dismissal of the affective fallacy, which entails "a confusion between the poem and its result" and which results from the attempt "to derive the standard of criticism from the psychological effects of the poem".[198] Patricia Waugh concedes that, within literary criticism, there is no value-free position from which a work can be evaluated. A work of literature can only be

[193] Patricia Waugh, "Value: Criticism, Canons and Evaluation", in: Patricia Waugh (ed.), *Literary Theory and Criticism: An Oxford Guide*, Oxford: Oxford University Press 2006, p. 70.

[194] Ian Buchanan, *A Dictionary of Critical Theory* (Oxford: Oxford University Press 2010), p. 382.

[195] Frank Raymond Leavis, *The Great Tradition: George Eliot, Henry James, Joseph Conrad* ([New York]: New York University Press 1963), p. 1.

[196] Terry Eagleton, *Literary Theory: An Introduction* (Minneapolis: University of Minnesota Press 1983), p. 10.

[197] Northrop Frye, *Anatomy of Criticism*, p. 18.

[198] W.K. Wimsatt, *The Verbal Icon: Studies in the Meaning of Poetry*, p. 21.

assessed in the light of a particular assumption about what constitutes literary quality, and these assumptions are inevitably particular to individual critical theories. Each conceptualisation of literature "already carries its own implicit value orientation".[199] The value of literary texts have sometimes been demonstrated using the "test of time" argument,[200] which suggests that works of a lesser quality are automatically winnowed out over the course of time. This argument is ultimately circular, however. It does not provide an explicit statement of the aesthetic qualities which have procured the continued interest, besides the endurance of the critical acclaim in itself.

The observation that evaluation cannot be based on stable and objective criteria ought not to lead to the conclusion that it is without relevance or importance, however. In literary studies, as perhaps in humanities research at large, the objective is rarely to provide a conclusive account of a text or to end a debate. As noted above, discussions about the quality of a literary work do not follow a progressive and cumulative programme, and the aim of a particular critical reading is usually to contribute to a discourse rather than to invalidate or to falsify earlier claims. Smallwood stresses that, although evaluative judgements cannot claim to be infallible, and although that they inexorably remain open to debate, evaluation and discrimination is inherent to the nature of criticism, as critics invariably pass judgements on the works they read.[201] A recognition of the fallibility and the situatedness of qualitative assessments might lead to a relativism in which all individual opinions are considered equal. As in the case for interpretations, however, evaluative judgements can be compared by considering the textual evidence that is used to support central arguments. Evaluation should not be based on biased or on fleeting impressions, as critics need to demonstrate that the qualities which are admired or disparaged are genuinely present in the text. Such explanations can enable peers to determine the accuracy and the propriety of the evaluation. To a large extent, the value of the close reading method also lies in the fact that it can enable scholars to collect the supportive evidence that they can ultimately use to buttress and to strengthen their central claims about literary works.

In this chapter, close reading has been defined as a method which concentrates on formal aspects. It is a non-reductive process in which scholars consider the full implications of the linguistic and literary features of a text. Close reading is not a mechanical process but one which is deeply responsive to the specificity of the literary work. As many of the activities which are central to the close reading are unpredictable and context-specific, they crucially resist formalisation. This capri-

[199] Patricia Waugh, "Value: Criticism, Canons and Evaluation", p. 74.

[200] Ibid., p. 73.

[201] Philip Smallwood, "Criticism, Valuation, and Useful Purpose", in: *New Literary History*, 28:4 (1 November 1997), p. 714.

cious and variable nature of close reading appears to be in conflict with the computer’s demand for explicit data and for predictable processes, setting pressing and compelling challenges for the very concept of algorithmic criticism.

Chapter 3

Current state of literary informatics

3.1. Introduction

Ever since computers were given the ability to process alphanumerical characters, scholars have experimented with the numerous ways in which the digital medium can query and manipulate works of literature. To a large extent, the appeal of computation lies in the unequalled speed with which information can be extracted from texts. Once the rules for the identification of features of interest have been implemented in an application, the actual time needed for the execution of this algorithm is, in most cases, negligible. Consequently, it becomes practicable to apply a set of algorithms iteratively to text bases of thousands or of millions of documents.[202] In addition, while the particular ways in which texts are processed by human scholars is often influenced, to a higher or lesser degree, by the scholar's mood or by levels of concentration, computers are incapable of fatigue, and apply the rules that are specified in an algorithm with unrelenting rigour and consistency to each text in the corpus.

This chapter discusses current scholarly practices in the field of literary informatics. As was discussed, many of its methods are based on text mining technologies. While text mining and NLP are rich and burgeoning areas of research, the aim of this chapter is not to describe the achievements and technical challenges in these areas as such. The scope remains limited to the tools and the concepts that have been developed or adopted by humanities scholars for the investigation or the clarification of literary texts. The main aim is to describe the concepts and principles that underlie text analysis tools, and to arrive at a better understanding of the methodology and the epistemology of computer-based literary research.

Studies in the field of literary informatics frequently make use of standardised text analysis tools. Many examples of such applications can be found via online directories such as Tapor[203] and Dirt.[204] The nature of literary informatics research can partly be examined by considering the basic functionalities which are offered by these tools. Evidently, the tools that have been developed in research projects have not all been made available publicly as independent production level services. This is not always possible, because tools have sometimes been developed for very

[202] Gregory Crane, “What Do You Do with a Million Books?”, in: *D-Lib Magazine*, 12:3 (2006).

[203] <http://www.tapor.ca/> (19 October 2013)

[204] <http://dirtdirectory.org/> (19 October 2013)

specific research goals, or for highly specialised data sets. The following section provides an overview of the general functionalities which are offered by these tools. The third section of this chapter focuses on the various ways in which data and tools have been used concretely within research projects, and on the various research questions that have been addressed using these tools.

3.2. Tools

3.2.1. Vocabulary

Many of the functionalities that can be offered by text analysis tools are based on counts of the words that appear within a text. Users can often upload texts in the plain text format, and the program can subsequently divide the text into smaller linguistic units, such as, for instance, its words or sentences. This preparatory process is generally referred to as "segmentation" or "tokenisation". Segmentation generally takes place on the basis of the spaces, punctuation marks and line breaks that occur in the text. Such notational conventions are currently used in virtually all of our written natural language texts, and they have been in use at least since the Carolingian Renaissance in the late 9th century. Scribes in the early Middle Ages introduced a number of rules aimed at rendering the ocean of words that was found in ancient *scriptura continua* in a more legible form. Innovations included the use of spaces in between words, the distinction between upper and lower case, and the insertion of punctuation to mark the end of a sentence.[205] On the basis of the notational conventions, which Feldman refers to as 'soft markup',[206] text mining applications can be developed for the recognition of units such as words, sentences or paragraphs. [207] The total number of words that are found are referred to as "tokens", and the unique words are called "types". Frequency lists, which count occurrences of types, form the basis for further statistical analyses.

If word segmentation takes place exclusively on the basis of the usage of spaces, this is arguably a rather crude method. Views may vary, for example, on what exactly constitutes a word. Brinton explains that there are several ways in which the boundaries of words may be determined. [208] He discusses a distinction between orthographic and semantic criteria. In the orthographic approach, a word is simply

[205] Paul Saenger, *Space between Words : The Origins of Silent Reading* (Stanford Calif.: Stanford University Press 1997), p. 10.

[206] Ronen Feldman, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, p. 3.

[207] Very few tools offer support for sentence segmentation. One tool which may be used is the Lingua::Eng::Sentence in the Perl programming language <http://search.cpan.org/~achimru/Lingua-Sentence-1.05/lib/Lingua/Sentence.pm> (10 October 2013)

[208] Laurel Brinton, *The Structure of Modern English a Linguistic Introduction* (Philadelphia: John Benjamins 2000), pp. 73–74.

defined as a string of characters delineated by spaces. Complications arise, however, in the case of hyphenated words, compound words or phrasal verbs. Compound words obviously form one semantic unit, but they are generally treated as separate types by tokenisation algorithms. John Kirk discusses an additional number of complexities. Scholars may have dissimilar views on whether words in different spellings, e.g. the difference between 'color' and 'colour' in British and American spelling, ought to be treated as different types. Dialect words "represent various pronunciation variants of the same lexical type".[209] Counts, furthermore, may be based either on the lemma, being the dictionary entry of the word, or on the variously inflected forms of words. One additional difficulty is that tools often ignore the distinction between characters in upper case and in lower case by default. In projects that study the religious nature of texts, for instance, the occurrence of capitals in masculine personal pronouns may be highly relevant. John Burrows also notes that words that are treated by the tokeniser as a single type may nevertheless have widely distinct meanings. Current methods seem unable to deal adequately with "truly polysemous words like "blue" where numerous literal meanings shade off into all sorts of metaphorical senses".[210] Since research projects are not always fully transparent with respect to the way in which they actually produce counts, Kirk stresses that the use of frequency lists invariably demands interpretation.[211] Frequency lists should routinely be treated with caution.

Word segmentation is complicated by the fact that the use of punctuation marks such as apostrophes and hyphens is generally unpredictable. The exact purpose of these characters often varies with the context and with personal stylistic preferences of authors. The identification of words and sentences in English texts is normally relatively easy for human readers who have a proficiency in that language, but, for computers, "dealing with hyphenated words, apostrophes, conventions of using single and double quotes, and so forth all require the programmer's attention".[212] Different text analysis tools may have implemented different rules for tokenising texts on the basis of spaces and punctuation marks, and such discrepancies evidently lead to different counts. Taporware, for instance, removes punctuation marks such as question marks, quotation marks and exclamation marks, but does not remove trailing and leading hyphens and asterisks. Voyant removes all punctuation marks, except for trailing quotation marks and hyphens. The Lexomics tool offers users the possibility to remove all punctuation, with the option, nevertheless, to retain hyphens or word-internal

[209] John M. Kirk, "Word Frequency: Use or Misuse?", in: Dawn Archer (ed.), *What's in a Word-List?: Investigating Word Frequency and Keyword Extraction*, Farnham: Ashgate 2009, pp. 19–20.

[210] John Burrows, "Never Say Always Again: Reflections on the Numbers Game", in: Willard McCarty (ed.), *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*, Cambridge: Open Book Publishers 2010, p. 26.

[211] John M. Kirk, "Word Frequency: Use or Misuse?", p. 33.

[212] Roger Bilisoly, *Practical Text Mining with Perl* (Hoboken, N.J.: Wiley 2008), p. 8.

apostrophes. While such differences may appear trivial, they have a direct impact on the counts of tokens, and, consequently, on the results of text analyses.

The resultant counts may be analysed in a variety of ways. A number of tools support the creation of a keyword in context list or a concordance. This entails, more specifically, a functionality in which occurrences of a given search term can be shown in combination with words that occur before and after this term. Users can usually specify the length of the fragments in such lists. Frequencies can also be shown as a distribution graph. Distribution refers to a process in which a full text is divided into segments, and in which information is provided about the frequency of a particular word within each of these segments. Tools may also offer a collocation display, in which a frequency list of the search term that was provided is shown in combination with frequency lists of the words that occur most frequently within a certain distance of the provided keyword. In co-occurrence searchers, text fragments can be retrieved in which two given words occur within a certain distance of each other. Users can generally indicate the number of other words that may occur in between these words. In all of these operations, the focus is exclusively on the frequencies of words. These operations fully disregard the original word order. Approach which ignore the original context of words are commonly described collectively as the “bag of words” model.[213]

3.2.2. Linguistic features

As literary writing typically employs a register of language that differs in a number of important ways from common colloquial language, systematic analyses of literary texts commonly focus on the distinctive ways in which narratives and ideas have been expressed. Analyses of the language can be based on manipulations of the unprocessed plain text, but the scope of textual analyses can be expanded considerably by employing software that can enrich the plain texts with detailed annotations about their linguistic characteristics. Data about the syntactic categories of words, for instance, can be produced by making use of part of speech (POS) taggers. Many of the taggers that are available today implement a method that was proposed originally by Eric Brill. His method entails a process in which syntactic categories are initially supplied by making use of a large lexicon which consists of words from the English language, together with their potential grammatical categories. Using this lexicon, the text can be tagged provisionally. The results of the first round of tagging are subsequently improved through a set of transformation rules, which can detect cases of incorrect categorisation. It can be stipulated, for instance, that certain combinations, such as a singular pronoun followed by plural verb, can never occur. Using such approaches, POS tagging can

[213] Roger Bilisoly, *Practical Text Mining with Perl*, p. 123. On the basis of the definition that is given, it must be decided that a KWIC list is not based on the bag of words model.

take place semi-automatically. If the default lexicon is not accurate for a specific corpus, the tagger needs to be trained. This means that users will need to tag a set of text manually, so that the application can generate a new lexicon and new transformation rules.[214] POS taggers have been developed by the Stanford NLP team[215] and within the OpenNLP project.[216] In the PERL programming language, the module Lingua::Eng:Tagger[217] can be used, and the NLP toolkit for Python similarly includes a POS tagger.[218] While POS taggers typically make use of algorithms and transformation rules, data on grammatical categories can also be provided solely by making use of a lexicon. One example is Docuscope.[219] It is a corpus, developed by Michael Witmore at the University of Wisconsin, of several million English words and phrases of words which have been associated manually with specific grammatical, semantic and rhetorical categories. The functional linguistic categories which are used in DocuScope are referred to as 'Language Action Types' (LATs). Docuscope matches the string in the central dictionary to one of the pre-defined LATs. The strings "I" and "me", for instance, are be labelled with the LAT 'FirstPerson'.

Many existing pre-trained POS taggers assume a regularised spelling and a predictable structure. They do not function accurately in all cases. Kaplan notes that the results of POS taggers are frequently unreliable when they are used to parse the often complicated syntax of poetry. Poems often have a syntax which is deliberately ambiguous, leading to a situation in which taggers can potentially assign multiple categories. In the case historical texts or texts in dialects, challenges are posed by the fact that spelling and syntax can vary along with different eras or different regions. Many of these challenges have been addressed in the *Metadata Offers More Knowledge* (MONK) project, which ran from 2007 to 2009 under the direction of John Unsworth. The aim of the project was to enable humanities scholars to engage in text mining on the basis of tools which were already in use within the field of corpus linguistics. As part of the MONK project, a vast data store has been created with approximately 2500 texts, which collectively contain more than 150 million words. The corpus covers texts from different geographic areas and different historical periods. The objective, nonetheless, was to let scholars query these texts in a uniform manner. To allow for such searches across dialects and across historical eras, all texts have been encoded using a specific variety of TEI-P5, which was referred to as TEI-Analytics. In this encoding,

[214] Eric Brill, "A simple rule-based part of speech tagger", in *Proceedings of the third conference on Applied natural language processing*, (Morristown, NJ, USA: Association for Computational Linguistics, 1992), p. 152.

[215] <http://nlp.stanford.edu/software/tagger.shtml> (28 May 2014)

[216] <https://opennlp.apache.org/> (28 May 2014)

[217] <http://search.cpan.org/dist/Lingua-EN-Tagger/Tagger.pm> (28 May 2014)

[218] <http://www.nltk.org/> (28 May 2014)

[219] <http://www.cmu.edu/hss/english/research/docuscope.html> (26 September 2013)

each individual word was connected to its lemma, and to its syntactic category. As part of the project, a new POS tagger, named Morphadorner, was developed.[220] Within the project, it was decided that "the tokenizer should not sunder what the typesetter has joined", and, following this logic, contracted forms such as "th'earth" or "nilt" were not expanded.[221] The expanded forms were provided, nevertheless, in the lemmatised version, which connects the word form that is found in a text to its dictionary form. The MONK datastore and the software that has been developed form important resources for literary scholars who seek to explore patterns and regularities in historical text collections, based on extensive linguistic data.

Next to grammatical categories, scholars may also be interested in using data on phonetic aspects. Words, when read aloud, obviously contain sounds, and literary devices are often based on a skilful use of such sounds. Data on phonetics are especially relevant for the analysis of poetry. Since the English language does not have a close correspondence between orthography and pronunciation, it is not possible to extract data about sounds directly.[222] Phonetic transcriptions may be produced, nevertheless, by making use of pronunciation dictionaries. This approach was followed in the development of the PoetryAnalyzer tool, which was created by David Kaplan at Princeton University.[223] Among other functionalities, PoetryAnalyser enables scholars to identify literary devices such as perfect rhyme and alliteration in a text. The detection of these features are based on prior phonetic transcriptions, which are made using the openly available Carnegie Mellon Pronunciation dictionary.[224] Other dictionaries are available as well, but, as Kaplan was mainly interested in poets from the United States, this specific resource was chosen because it offers data on the pronunciation of American English. The tool can transcribe the tokens found in the texts via a lookup in the dictionary. Kaplan conceded that such direct lookups produce a small margin of errors, since certain words in the English language, such as "record" or "minute", may have different pronunciations, depending on their syntactic function. No measures were taken to correct these errors, however. Texts may also contain proper nouns such as personal names and geographical terms, and these will in most cases not be listed in the dictionary. In addition, the method cannot deal adequately with diachronic and synchronic variations in pronunciation. The PoetryAnalyzer software is less effective for investigating texts by British poets, for instance.

[220] <http://morphadorner.northwestern.edu/>

[221] John Unsworth & Martin Mueller, *The MONK Project Final Report*, (2009), p. 6.

[222] Susan Hockey, *Electronic Texts in the Humanities: Principles and Practice*, p. 78.

[223] David Kaplan & D.M. Blei, "A Computational Approach to Style in American Poetry", in *Seventh IEEE International Conference on Data Mining*, (2007), pp. 553–558.

[224] <http://www.speech.cs.cmu.edu/cgi-bin/cmudict> (15 March 2013)

Phonetic transcriptions can alternatively be produced using text-to-speech software, of which MaryTTS is doubtlessly the most significant example.[225]

3.2.3. Semantic contents

The quantitative and statistical approach towards studying literature has frequently been under attack as a result of its perceived superficiality. Indeed, if the concern is predominantly with the symbols that carry meaning, rather than with the meaning itself, this is admittedly a rather shallow form of engagement. In this mode of textual analysis, the relationship to literary works appears to be equivalent to the manner in which texts are treated in disciplines such as bibliography, codicology, book history or library studies. While these latter fields take books or other carriers of information as their primary research objects, they do not necessarily concern themselves with the contents of these texts. These disciplines focus on data about texts, and may study the reception, the production processes, or the physical appearance of items. Studies in the field of algorithmic criticism, by contrast, aim to contribute to an improved understanding of the contents and of the more profound thematic concerns of literary works, and various attempts have been made to bridge the gap between the lexical codes and their semantic contents.

One of the ways in which the semantic aspects of texts may be uncovered is by making use of lexicons which map the text's tokens to pre-defined semantic categories. Examples of applications in which this principle is implemented include the Harvard General Inquirer,[226] the Linguistic Inquiry and Word Count (LIWC) tool,[227] the UCREL Semantic Analysis System (USAS)[228] and DocuScope.[229] The Harvard General Inquirer, firstly, consists of 182 categories, each of which are connected to an extensive list of words. The category "negative", for instance, contains over 2290 entries. The DocuScope tool, which was also mentioned in a previous section, can provide data both about grammatical features and about semantic aspects. Phrases such as "whilst," "when he," "as he", for example, are labelled with the LAT "Narrative Time". LIWC, thirdly, consists of general pre-defined semantic categories for the words that are used in a text. It uses categories for positive or negative emotions, mental processes, self-references, and causal words. The tool can therefore be used effectively for sentiment analysis. Fourthly, the UCREL Semantic Analysis System (USAS) application, which was developed at

[225] MaryTTS is as "an open-source, multilingual Text-to-Speech Synthesis platform written in Java", "originally developed as a collaborative project of DFKI's Language Technology Lab and the Institute of Phonetics at Saarland University". It offers support for "German, British and American English, French, Italian, Swedish, Russian, Turkish, and Telugu". <http://mary.dfki.de/> (12 June 2013)

[226] <http://www.wjh.harvard.edu/~inquirer/> (12 June 2013)

[227] <http://www.liwc.net/> (12 June 2013)

[228] < http://ucrel.lancs.ac.uk/usas/> (12 June 2013)

[229] <http://www.cmu.edu/hss/english/research/docuscope.html> (12 June 2013)

the University of Lancaster, consists of 21 major domains, which expand into 232 more specific semantic field tags.

Next to such lexicon-based approaches, investigations of the semantic contents of texts can also be based on statistical processing of the vocabulary. Topic modelling is the prime example of this approach. It is a generic term which refers to a range of algorithms that can be used to determine the topics that occur in a text on the basis of the vocabulary used in individual documents. Topic modelling is performed most frequently on the basis of an algorithm which is known as Latent Dirichlet Allocation (LDA), which was first discussed in an article by David Blei et al.[230] LDA is also implemented in MALLET, a Java-based tool created at the University of Massachusetts-Amherst.[231] The tool takes a text collection as input, and produces a number of topics as a result. The topics that are returned by MALLET concretely consist of unnamed lists of words. The number of topics to be returned have to be supplied as a parameter before running the algorithm. Users of MALLET need to inspect and interpret the lists and provide topic labels themselves. The central idea of Topic Modelling is that documents contain topics, and that these topics manifest themselves through specific words. If certain words co-occur frequently in the same documents, in ways that have been defined as statistically significant, these are assumed to be about the same topic. MALLET does not only return the topics, but also compiles a list of the documents containing these various topics. In this way, Topic Modelling can also be used to cluster the documents that focus on specific topics.

3.2.4. Data analysis

Studies in the field of literary informatics concentrate to a large extent on a description of the style of literary texts. The term “style” is used ubiquitously within literary and linguistics research, and, as a result, it is open to many different interpretations and definitions. Based on an elaborate survey of the various conceptualisations of the term within German, French and Dutch traditions in linguistic and literary scholarship, Herrmann, van Dalen-Oskam and Schöch define the term “style” broadly as “a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively”.[232] This conceptualization of “style” has also been adopted in this thesis. The formal features which are mentioned in the definition may refer to “linguistic features at the level of characters, lexicon, syntax, semantics”, and also to “features going

[230] David M. Blei, Andrew Y. Ng & Michael I. Jordan, “Latent Dirichlet Allocation”, in: *The Journal of Machine Learning Research*, 3 (1 March 2003).

[231] <http://mallet.cs.umass.edu/> (2 March 2014)

[232] J. Berenike Herrmann, Karina van Dalen-Oskam & Christof Schöch, “Revisiting Style, a Key Concept in Literary Studies”, in: *Journal of Literary Theory*, 9:1 (2015), p. 44.

beyond the sentence, such as narrative perspective or textual macro-structure".[233] The authors argue that a text's style can be characterised via a careful examination of all occurrences of these formal features, and of the complicated ways in which these features can be combined.

On an abstract level, computational literary analyses begin with a quantification of some of the formal features which constitute the style of a text, making use of the applications which have been discussed in the previous sections. The process of quantification results in a series of variables and of associated values for these variables. Without further processing, it is usually difficult to see patterns within the frequencies of such style markers. To understand the nature of data sets more fully, it is often necessary to apply additional techniques. The statistical procedures which can help to model and to clarify existing data collections are commonly referred to as "statistical learning" techniques or "machine learning" techniques. James et al. make a distinction between supervised techniques and unsupervised techniques.[234] Examples of supervised techniques include Naive Bayes classification, logistic regression and Support Vector Machines.[235] Supervised machine learning techniques can be used, among other purposes, to classify texts. In the case of classification, researchers firstly need to assign labels or categories to the texts in a training set. Using these labels, the classification algorithms can construct a statistical model which may be used to make predictions about the categories of new and unlabelled texts. James et al. make an interesting distinction between "prediction" and "inference". In the former approach, statistical learning techniques are applied primarily to classify unlabelled texts. In the case of inference, the focus lies mainly on an examination of the formal properties which, according to the model that was created, are typical for the texts in specific categories.

In the case of unsupervised learning techniques, scholars do not supply prior information about the potential categories of texts. This second class of techniques aims to discover patterns, clusters or relationships within unlabelled data. Patterns can, in some cases, be found simply by sorting or filtering the data on the basis of a particular data value. A widely used and more advanced unsupervised machine learning technique is Principal Component Analysis (PCA). It is a form of multivariate analysis, which reduces the complexity of a multidimensional data set through the creation of a number of new composite variables which account for most of the variability of the original variables. These new variables are referred to as the principal components. By plotting a limited number of principal components, certain patterns can be explored in the global distribution of the data values.

[233] J. Berenike Herrmann, Karina van Dalen-Oskam & Christof Schöch, "Revisiting Style, a Key Concept in Literary Studies", p. 44.

[234] G. James et al., *An Introduction to Statistical Learning, with Applications in R* (Springer 2013), p. 1.

[235] Joachim Diederich (ed.), *Rule Extraction from Support Vector Machines* (Berlin: Springer 2008).

Diagrams in which PCA are visualised can disclose the words that occur in similar frequencies, or can indicate the texts which use very similar words. Differences between texts can additionally be characterised via the calculation of the Euclidean distance or of the cosine similarity. The distances between texts may be clarified in the form of a dendrogram. In such diagrams, the texts which are most similar form a single branch, and texts which display fewer similarities do not form a union until a much later stage. As such, the method provides a highly intuitive method for clarifying the differences and the similarities between texts.

In general, quantitative analyses of data about the stylistic properties of text can only clarify patterns of differences and similarities. In literary informatics research, such statistical comparisons are eventually used to answer questions which are more directly germane to literary criticism. As will be discussed in the next sections, the results of such procedures can be used, for instance, to compare texts written by different authors, speeches uttered by different literary characters, texts written by male and by female authors or texts from different literary genres or periods.

3.3. Studies

3.3.1. Methodology

This section concentrates on the concrete research questions that have been addressed using text analysis tools. To characterise current practices, a number of representative or exemplary studies have been analysed. As a first step, an inventory was made of the practical studies which are discussed in the various contributions to the *Blackwell Companion to Digital Humanities* and the *Blackwell Companion to Digital Literary Studies*.[236] This initial list was extended by identifying all articles which discuss computer-based practical work, published either in *Literary and Linguistic Computing*, the *Digital Humanities Quarterly* or the *Journal of Digital Humanities* during the period in between 2009 and 2014. In the following sections, these studies have been clustered by considering the main literary phenomena they concentrate on.

3.3.2. Literary genres

Quantitative methods have often been used to study the stylistic differences between literary genres. Many of the studies in this category are based on analyses of word frequencies. Hugh Craig, for instance, has conducted a study of 25 Shakespeare plays, based of counts of the 12 most common words. The objective of

[236] *A Companion to Digital Humanities* (Malden, MA: Blackwell 2004) and Susan Schreibman, Ray Siemens, & John Unsworth (eds.), *A Companion to Digital Literary Studies* (Malden, MA: Blackwell 2007).

the study was to investigate if the groups that could be produced through statistical analyses of word frequencies correspond in some way to traditional divisions into comedies, tragedies and history plays. In Craig's study, a PCA revealed that the history plays indeed use a different set of words. Furthermore, there is a basic distinction between plays which contain a high frequency of the pronouns "you" and "I", on the one hand, and plays in which "of", "and" and "the" are commonly used. Craig argues that this reflects a division within the corpus between texts which contain "interactive dialogue" and plays which mostly contain "description and narration".[237]

Literary research on the differences between genres can alternatively be based on data about syntactic categories. This potential has been explored by some of the studies that have been conducted under the aegis of the Stanford Literary Lab. This lab was founded in 2010 by Matthew Jockers and Franco Moretti, and "discusses, designs, and pursues literary research of a digital and quantitative nature".[238] The first pamphlet of the Literary Lab addresses the question whether literary genres, such as the gothic novel or the Bildungsroman, can be recognised by computer algorithms. To investigate this issue, a range of experiments were carried out on British novels from the 19th century taken from the Chadwick-Healey collection.[239] The experiments focussed on two sets of data. The first set of data contained the LATS produced by DocuScope. The research was also based on information on the most frequent words. The data sets were analysed using principal component analysis and clustering technologies.[240] A central aim was to discover if certain combinations of LATs or frequent words are also characteristic of specific genres. The results were tested against genre assignments from existing bibliographies. Results suggested, however, that the techniques were best at recognising authorship, rather than genre. There were some notable differences between texts from different historical periods, but for genres that flourished during the same historical periods, the results were poor.[241]

The same topic was revisited in the fifth pamphlet of the literary lab. The pamphlet investigates the hypothesis that literary genres can be characterised through their use of specific grammatical constructions. It was also assumed that Gothic novels often contain fixed combinations of articles, nouns and prepositititions, as in phrases like "the Castle of Otranto", or "the Rock of Glotzden".[242] To

[237] Hugh Craig, "Stylistic Analysis and Authorship Studies", in: *A Companion to Digital Humanities*, Oxford: Blackwell 2002, pp. 274-277.

[238] <http://litlab.stanford.edu/> (4 August 2013)

[239] <http://collections.chadwyck.co.uk/> (4 August 2013)

[240] More specifically, the "dist" and "hclust" functions available in the open-source "R" statistics application were used.

[241] Sarah Allison et al., *Quantitative Formalism: An Experiment*, (Stanford: Stanford Literary Lab 2011).

[242] Ryan Heuser & Long Le-Khac, *A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method*, (Stanford Literary Lab 2012), p. 2.

investigate these hypotheses, British novels were selected from the Chadwyck-Healey nineteenth-century database. From these novels, the sentences were isolated, and counts were created of particular types of sentences. Among other types, distinctions were made between sentences that consist of a single independent clause, sentences which contain an independent clause followed by a dependent clause, and sentences in which an independent clause is followed by a non-finite clause. In the case of sentences with multiple clauses, different types of conjunctions were also analysed. It was found, among other things, that Charles Dickens and Ann Radcliffe predominantly use sequencing conjunctions, in sentences that have a dependent clause preceded by an independent clause, and that Walter Scott mostly uses relative or "defining" dependent clauses. However, the findings were difficult to generalise into conclusive statements about correlations between grammatical constructions and literary genres.[243]

3.3.3. Literary characters

Stephen Ramsey has conducted a study which focused on Virginia Woolf's novel *The Waves*. The novel contains six related monologues, each spoken by distinct characters, who together narrate a related series of events. Ramsay has explored the differences between these six monologues on the basis of the term frequency-inverse document frequency (tf-idf) formula. In regular word lists, frequencies are usually distributed according to Zipf's Law, which states that there are normally small numbers of words that occur very frequently, and large numbers of hapax legomena, which are words that occur only once. The tf-idf formula assigns weights to the bare counts of the words, which are calculated by dividing the regular frequency of the type by the total number of texts that contain the type. Consequently, they have a higher value when words are infrequent. Ramsey demonstrates that the word lists that are generated in this way can indeed be used to disclose some of the central differences between the six protagonists. The list for the Australian character Louis, for instance, convincingly exposes a consciousness of his accent and of his nationality. Ramsey explains that statistical processing typically results in a paratext that can be interpreted and explained by the scholar, and that such resources may help to "to confirm or deny the 'serendipitous reading' of literary critics".[244]

Kyle Mahowald, at the Language Lab at MIT, has investigated occurrences of y- and th- pronouns as used by characters in Shakespeare's plays.[245] Making use of the Natural Language Processing toolkit in Python, a mechanism was developed to

[243] Ryan Heuser & Long Le-Khac, *A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method*, p. 10 and passim.

[244] Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism*, p. 14.

[245] K. Mahowald, "A Naive Bayes Classifier for Shakespeare's Second-Person Pronoun", in: *Literary and Linguistic Computing*, 27:1 (10 November 2011).

identify all occurrences of second person pronouns. In addition, a Naïve Bayes classifier was built to predict whether these pronouns were th- or y-pronouns, by making use of collocations. It was found that words such as “lordship”, “madam” and “sir” were most useful in classifying a pronoun as a y-pronoun, suggesting that these are mainly used to address a personal with a higher social status. Unsurprisingly, it was also found that th-pronouns were used mostly by characters with a higher status to address persons from lower social classes.

3.3.4. Date of creation

The studies that have been discussed so far are critically based on word segmentation and on analyses of frequency lists. Studies may also be based on random substrings. Richard Forsyth has investigated the differences between the early and the late poetry of W.B. Yeats using a technique which is called Monte Carlo Feature Finding. [246] The main aim of the study was to develop a method for dating texts. In this method, substrings are extracted in a manner that is fully opaque to the occurrence of word boundaries. Each substring is ranked according to its distinctiveness, which was measured on the basis of Chi-squared. The study suggested that short random substrings can indeed be used to classify texts either as an early or as a late poem. The methodology used in this particular study seems deeply remote from that of traditional literary studies. Substrings are, in most cases, entirely meaningless. While the algorithms enjoyed a degree of success in categorising poems, the patterns that were created are evidently difficult to interpret for literary critics. As such, the study fails to contribute to an understanding of how the early and the late poems differ precisely.

3.3.5. Authorship attribution

Word frequency lists have also been used successfully in studies that aim to suggest a probable author for texts whose authorship is disputed. Various studies have shown that patterns in the usage of vocabulary are strongly distinctive for individual authors, and that, as such, frequency lists can serve as an author's individual fingerprint. Authorship attribution studies generally make a distinction between lexical words and function words. The first term is used to refer to words such as nouns and verbs which usually carry most of the meaning in a sentence. Lexical words are often selected consciously, and the exact denotation or connotation usually varies strongly along with a word's context. Function words, by contrast, mostly have a relatively stable meaning. The term comprises words such as pronouns, articles and prepositions, which are assumed to be chosen unconsciously. Studies in the field of authorship attribution predominantly focus

[246] Richard S Forsyth, “Stylochronometry with Substrings, or: A Poet Young and Old”, in: *Literary and Linguistic Computing*, 14:4 (1999).

on function words. They make use of a "base strata of language where imitation or deliberate variation can be ruled out".[247]

Authorship attribution studies usually demand a meticulous preparation of the source materials, since such investigations largely exploit the distributions of function words which are "especially resistant to intentional authorial manipulation".[248] Burrows notes that, in the case of older texts, it can be useful to standardise the spelling or to expand contracted forms.[249] In addition, certain homographic words can also be tagged so that the different grammatical uses of words can be distinguished. Hoover has found that the accuracy of attribution tests improves when proper nouns, inflected words, personal pronouns are removed from the corpus. Hoover also proposes a removal of all dialogue from novels, but since dialogue is not always rendered typographically distinct, this often requires a degree of interpretation.

A large number of authorship attribution studies have made use of the delta method, which was developed originally by John Burrows. The method assumes a corpus which contains a work whose author is unknown, together with works by a number of potential authors. The delta value of a work, or a group of works, can be calculated by considering the difference between "the z-scores for a set of word-variables in a given text group and the z-scores for the same set of word-variables in a target set".[250] The delta value is the mean of the absolute values of these differences. Importantly, delta ignores the difference between positive and negative values. The central assumption in Burrow's method is that the probable author of the unassigned work can be found by comparing the delta value for this individual work to the values for the potential authors. If the delta is low, there is a higher probability that the works are by the same author. In his article *Never Say Always Again: Reflections on the Numbers Game*, Burrows discusses a number of exemplary authorship attribution studies which are based on delta. The first of these concerns the novel *St. Ives*, which was begun by Robert Louis Stevenson, and which was completed by Arthur Quiller Couch after Stevenson's death in 1894. For the study, Burrows took 72,000-word samples from the novels and stories of Stevenson and Quiller Couch, together with 12,000-word samples from authors who were active during roughly the same period. The tests pointed to Stevenson as

[247] Hugh Craig, "Stylistic Analysis and Authorship Studies", p. 273.

[248] David Hoover, "Word Frequency, Statistical Stylistics and Authorship Attribution", in: Dawn Archer (ed.), *What's in a Word-List?: Investigating Word Frequency and Keyword Extraction*, Farnham: Ashgate 2009, p. 35.

[249] John Burrows, "Textual Analysis", in: Susan Schreibman, Ray Siemens, & John Unsworth (eds.), *A Companion to Digital Humanities*, Oxford: Blackwell 2002, p. 269.

[250] J. Burrows, "Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship", in: *Literary and Linguistic Computing*, 17:3 (1 September 2002), p. 271.

the author of the first 29 chapters, and to Quiller Couch for the authorship or the remaining 11 chapters, which is in line with the available historical evidence.[251]

3.3.6. Themes

Analyses of the words that are used within a corpus can also be applied to find the words associated with specific themes or specific topics. Martha Nell Smith has explored the use of erotic language in the correspondence of Emily Dickinson, on the basis of a corpus of about 200 letters that were exchanged between the poet Emily Dickinson and her sister-in-law Susan Huntington Dickinson. As part of the study, an application was built in which scholars could classify the degree of eroticism, on a scale of 0 to 5. The letters that were rated manually formed the training material for the classifier. The application attempted to classify the remaining texts, using a method based on Naieve Bayesian logic. The application also indicated the words that were found to be associated with eroticism. One of the surprising outcomes was that the word "mine" emerged as a marker of eroticism. According to Smith, certain words only assume an erotic subtext because of their co-occurrence with other words, and computers are better at exposing such co-occurrences because human readers "may unselfconsciously divide epistolary subjects within the same letter [...] into completely separate categories".[252] The complete disregard for rhetorical structure may help to detect connections which may previously have escaped scholars.

Kao and Jurafsky have investigated the literary quality of poems, by comparing poetry written by skilled professional poets to texts written by amateur poets.[253] Assuming that the first class of poems is of a higher literary quality than the poetry in the latter class, Kao and Jurafsky examined the differences between these two sets of poems through a quantification of some of the linguistic and semantic features. The latter data were obtained by making use of HGI and LIWC. The counts obtained via these semantic taggers were normalised for the length of the poem. To investigate the differences between professional and amateur poetry, a logistic regression model for all 16 metrics was implemented in the R package. It was found that, while professional poets are significantly less likely to use words that explicitly refer to negative emotions than amateur poets, the usage of words with negative connotations was found to be roughly the same. This suggests that poets mainly evoke sentiments through connotations rather than through direct

[251] John Burrows, "Never Say Always Again: Reflections on the Numbers Game".

[252] Martha Nell Smith et al., ""Undiscovered Public Knowledge": Mining for Patterns of Erotic Language in Emily Dickinson's Correspondence with Susan Huntington (Gilbert) Dickinson", in *Digital Humanities*, (2006), pp. 252–255, p. 254.

[253] Justine Kao & Dan Jurafsky, "A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry", in *NAACL Workshop on Computational Linguistics for Literature*, (2012), pp. 8–7.

denotation. In addition, the results also indicated that professional poets tend to use concrete imagery rather than references to abstract or general concepts.

Archer, Culpepper and Rayson have used a related method to investigate “key domains” in Shakespeare’s drama.[254] The authors have investigated the ‘aboutness’ of six comedies and tragedies using USAS. The semantic categories that were found enabled the authors to investigate the images that served as the tenors in metaphors that were used in the love comedies and in the love tragedies. The differences between the two sets of poems were compared by calculating the log-likelihood from the frequencies of the semantic tags that were identified. Through this procedure, the most overused and the most underused semantic fields could be identified for both groups of plays. It was found, among other things, that the categories ‘liking’ and ‘intimate/sexual relationships’ was underused in the love tragedies, in comparison to the love comedies. The tragedies had a higher occurrence of tags such as ‘war’ and ‘religion and the supernatural’.

Lisa Rhody has used Topic Modelling to investigate figurative language in a genre of poetry called Ekphrasis, which entails “poems written to, for, or about the visual arts”.[255] Rhody used LDA to identify topics in ca. 4,500 poems, and encountered a large number of difficulties. An obvious complication was that, while Topic Modelling assumes that there is a close connection between the words that are used and the topics they refer to, figurative language obviously uses word senses in ways that differ widely from their conventional meanings. Furthermore, ekphrastic poetry typically discusses works of art which depict certain phenomena. The phenomena that are shown in paintings or in drawings, however, do not necessarily constitute the central topics of the poem. A number of topics veered felicitously towards specific images and specific themes, but MALLET also returned various lists of words which seemed to have little in common. Such “semantically opaque” topics are frequently of particular interest for the literary critic, however, as they may reveal particular forms of discourse. They can also indicate words that are typical for specific genres or for specific themes. Rhody’s study has led to an improved understanding of the words that are commonly used in elegiac poetry, for instance. Furthermore, lists of words which seem to have little in common may stimulate scholars to revisit the poems in which these words occur, and such more directed forms of close reading can show that these texts share particular thematic concerns. Rhody argues that working with MALLET can help scholars to make new discoveries, “not because topic modelling works

[254] Dawn Archer, Jonathan Culpeper & Paul Rayson, “Love – “a Familiar or a Devil”? An Exploration of Key Domains in Shakespeare’s Comedies and Tragedies”, in: Dawn Archer (ed.), *What’s in a Word-List?: Investigating Word Frequency and Keyword Extraction*, Farnham: Ashgate 2009.

[255] Lisa M. Rhody, “Topic Modeling and Figurative Language”, in: *Journal of Digital Humanities*, 2:1 (2012).

perfectly, but because poetry causes it to fail in ways that are potentially productive for literary scholars”.[256]

3.3.7. Lexical repetitions

When a text has been tokenised into words, it is also possible to study repetitions of words or groups of words in a text. In a pioneering study, Tanya Clement has applied several algorithms for the exploration of repetitive patterns in the novel *The Making of America* by Gertrude Stein, a postmodern work which in itself is “almost impossible to read [...] in a traditional, linear manner”.[257] Software has been developed which divided the text into so-called n-grams, or patterns of co-occurring words. The study focused, more specifically, on series of three consecutive words. It was shown that the exact same trigram re-occurred in different sections of the novel. Moreover, in a number of passages, verbatim repetitions of even longer units were found. Such repetitions are difficult to identify without a computer, since full-text searches for specific repeated phrases assumes a “pre-knowledge that they exist—a nontrivial feat in the midst of the more pervasive and shorter repetitions that make up each section”.[258] As part of the study, a number of visualisation tools have also been devised to indicate how repetitive patterns are distributed over the various sections of the novel. Clement’s analyses suggested that the placement of repetitions indeed follows a highly regular pattern. While many studies in the field of literary informatics study vocabulary under the bag of words model, in which the original context of words is lost, this approach is evidently unsuitable for studies that focus on the structural repetition of words and phrases, as information about the original location of the words is generally relevant for studying the nature and the distribution of such echoes. In such studies, a bag of words model is inappropriate.

3.3.8. Rhyme and meter

Various scholars have used the computer to support of prosodic analyses. Malcolm Hayward, an English scholar at Indiana University of Pennsylvania, has conducted an exemplary study which focussed on 1000 lines of poetry, consisting, more specifically, of ten 100-line samples from ten different authors, including John Donne, Alexander Pope, John Keats and Robert Browning. All lines in the corpus were iambic pentameters. Hayward’s aim was to “provide a quantitative basis for

[256] Lisa M. Rhody, “Topic Modeling and Figurative Language”, n.pag.

[257] T. E. Clement, ““A Thing Not Beginning and Not Ending”: Using Digital Tools to Distant-Read Gertrude Stein’s The Making of Americans”, in: *Literary and Linguistic Computing*, 23:3 (5 September 2008), p. 361.

[258] Ibid., p. 363.

comparisons between poets",[259] and to investigate the metrical variations in these poetic lines, using a connectionist model of poetic meter. The connectionist model is based on the idea that each syllable in the line "is connected to five other units, representing possible inputs towards stress from intonation, lexical features, prosody, syntax, and interpretation". Hayward has manually assigned scores for each of these dimensions. At the lexical level, for example, primary stress was marked with a '2', secondary stress with a '1', and all unstressed syllables received '0'. The scores were analysed using statistical software, and this resulted in "a measurement of the potential activation of metrical stress for each of the ten positions for that particular line of poetry". Next, multivariate analyses were performed for the stress of each individual syllable, and these revealed significant differences among all ten poets represented in the study. From these findings, the conclusion was drawn that poets indeed display a highly idiosyncratic behaviour with respect to metrical variation.

Experiments with automated phonetic transcriptions have been conducted by, amongst others, David Kaplan. Kaplan was interested in the question if differences and similarities between poems can also be displayed visually. To explore this question, Kaplan developed a series of algorithms for the quantification of specific features of texts written by Northern-American poets.[260] Kaplan has also developed a number of algorithms that could use these transcriptions, together with data on syntactic categories produced by a POS Tagger, to produce a total of 84 different metric values for each poem. In the study, data were produced about occurrences of alliteration, assonance, consonance, perfect rhyme, slant rhyme and half rhyme, amongst other aspects. Automating the scansion is complicated by the fact that the placement of stresses depends, to a large degree, upon the meaning of the line. Using Principal Component Analysis and metric CMDS, these multiple values were visualised in a two-dimensional plot. The software developed "showed an ability to distinguish poetry texts based on a combination of salient features not traditionally used in computational prose text analysis but traditionally relied upon for poetry analysis".[261]

3.3.9. Allusions

Data on grammatical categories and on lemmas can also be applied usefully in explorations of literary allusions. Walter Crane explains that allusion can be viewed as a reference of one text fragment to another text fragment, and that they can be

[259] Malcolm Hayward, "Analysis of a Corpus of Poetry by a Connectionist Model of Poetic Meter", in: *Poetics*, 24:1 (July 1996), n.pag.

[260] David Kaplan & D.M. Blei, "A Computational Approach to Style in American Poetry".

[261] David Maxwell Kaplan, *Computational Analysis and Visualized Comparison of Style in American Poetry* (Princeton University 2006), p. 31.

either direct or indirect.[262] Direct allusions, firstly, are verbatim repetitions. The difference with quotations is that imitative textual allusions also invoke the context of the original, and that a knowledge of the reference is needed to interpret the allusion. Indirect allusions are essentially reworked versions of the original text, and they are, for this reason, more difficult to recognise algorithmically. Crane distinguishes a number of different types. Two fragments may contain identical words in a different order. Two passages may also share lemmas. Additionally, there may be a syntactical identity, a metrical identity and a semantic identity. Coffee et al. have noted that while computers may be used to identify such textual parallels, the automated detection of allusions is complicated by the fact that textual parallels are not always of literary significance. To distinguish meaningful allusions from other forms of parallels, the researchers have devised a model using data supplied by human critics. Parallels based on an "expanded feature set including bi-gram frequency, frequency of individual words, character-level n-grams and edit distances" were analysed through "support vector machines" and "random forests", and these experiments resulted in a number of statistical rules to identify the characteristics of meaningful allusions.[263] The algorithms have also been used to explore allusions to Vergil's *Aeneid* in the first book of Lucan's *Civil War*. The tool produced over 2,500 textual parallels, but many of these were found to be irrelevant. Nevertheless, the results that have been generated by Tesserae enables the developers to train and to further refine the algorithms.[264]

3.4. Discussion

As was discussed, computers can produce data about a broad range of textual aspects. On the basis of the spaces that are used in between words, computer applications can produce counts of the total number of words and of the total number of unique words. Tools in the field of NLP can additionally be used to produce data about the syntactic or grammatical categories of words, their lemmatised forms, their pronunciation, and their signification. Once they have been collected, these multifarious data can be analysed and visualised in a myriad of ways. In recent decades, quantitative approaches have been applied to study the formal aspects of literary genres, characters, themes, allusions and literary themes, among many other topics.

Notably, many of the studies which have been surveyed in this chapter have focused primarily on the further development and refinement of the methodology,

[262] Gregory Crane & David Bamman, "The Logic and Discovery of Textual Allusion", in: *ACL Language Technology for Cultural Heritage*, (2008).

[263] Neil Coffee et al., "Modelling the Interpretation of Literary Allusion with Machine Learning Techniques", in *Digital Humanities 2013*, (Nebraska–Lincoln: 2013), n.pag.

[264] N. Coffee et al., "The Tesserae Project: Intertextual Analysis of Latin Poetry", in: *Literary and Linguistic Computing*, 28:2 (20 July 2012).

rather than on the creation of new knowledge about literary texts. In *Testing Burrows's Delta*, for instance, David Hoover assesses the effectiveness and the accuracy of the delta method, and, using a sample corpus in which authors are all known already, demonstrates that the accuracy increases when a larger number of words are taken into account, or when personal pronouns are removed.[265] In a comparable study, Eder aimed to quantify the impact of textual errors on the accuracy of authorship attribution methods. The researchers had introduced varying numbers of typing errors in text corpora, in order to determine which level of noise would be acceptable.[266] The objective of the studies that were carried out for the first pamphlet of the Stanford Literary Lab was similarly to develop a procedure that can be used to create the exact same clusters and classifications that have been produced earlier by human scholars.[267] The assumption is that when the algorithms work correctly for a sample corpus of a modest size, the same rule-based approach can also be followed to classify texts which had previously been neglected.

As the current methodology is still of a probationary and experimental nature, studies which aim to calibrate or to meliorate the toolset are very beneficial. The ultimate aim of literary informatics research, however, is to concoct and to implement analytic methods which can genuinely advance the emanation of new ideas and new insights. Scholars ought to explore the ends to which these instruments can be put, and they need to evaluate the relevance or the value of these new methods, by enlisting these in the service of broader humanistic questions. Algorithmic criticism initially converts works of literature into numbers, but, ultimately, these numbers need to be converted in turn into qualitative or interpretative statements which can challenge, confirm or enrich our understanding of the texts that have been quantified.

[265] D. L. Hoover, "Testing Burrows's Delta", in: *Literary and Linguistic Computing*, 19:4 (1 November 2004).

[266] M. Eder, "Does Size Matter? Authorship Attribution, Small Samples, Big Problem", in: *Literary and Linguistic Computing*, (14 November 2013).

[267] Sarah Allison et al., *Quantitative Formalism: An Experiment*.

Chapter 4

Research data of literary informatics

4.1. Introduction

Literary informatics aims to fuse the scholarly objectives of literary studies with the affordances of digitisation, and concentrates, additionally, on the epistemological consequences of adopting computational approaches. In a minimalist scenario, the new methods are used principally to serve existing objectives and to rationalise established heuristics. David Berry stresses, nevertheless, that digital methods "have profound effects on all aspects of the disciplines".[268] Computation is undeniably characterised by a particular logic, and the use of technology demands a full or a partly subjection to this logic. Computer-based scholarship is initially a hybrid form of scholarship, in which parts of the conventional aims and methods are merged with the modus operandi that follows from technical exigencies. The precise outcome of the encounter between the digital and the traditional are often difficult to predict. It seems clear, nevertheless, that a careful examination both of the nature of digital methods and of the needs of the scholarly field in which these methods are adopted can help to identify potential incompatibilities, as well as opportunities for a productive confluence.

The two critical methods which are contrasted in this thesis have been discussed, to some extent, in the previous two chapters. Chapter 2 of this thesis has described conventional approaches to studying poetry, and Chapters 3 has explained some of the ways in which literary texts can be analysed via digital means. Chapter 3 has characterised the field of literary informatics in practical terms, concentrating on concrete tools and on specific research projects. This chapter seeks to characterise the nature of literary informatics research on a more conceptual level. One crucial characteristic of computer-based scholarship in general is that it is based on data. Digital humanities research often begins with a process in which the artefacts that are studied are converted into discrete data values, and the eventual scholarly claims are commonly based on statistical analyses of these data sets. Despite the general importance of the concept of research data, which Christine Borgman has referred to as "the foundation of scholarship",[269] it can be observed that there is still a degree of uncertainty about

[268] David M. Berry, "Introduction", in: *Understanding Digital Humanities*, New York: Palgrave Macmillan 2012, p. 13.

[269] Christine Borgman, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*, p. 115.

the precise meaning of the term within humanistic research.[270] It may have different meanings within the various branches of the humanities, and it may confusingly refer to different types of scholarly artefacts. This chapter introduces a number of terms that can be used to describe generic aspects of research data, and provides a definition of the various data types that are used and created by scholars. This framework is subsequently used to characterise the data that are produced by scholars in the field of literary informatics.

4.2. Definitions of research data

In recent years, there has been a growing interest in the potential advantages of the curation of research data among researchers, politicians and funding agencies, and whereas data, as noted, is a broad and a convoluted term, a number of generic definitions have been proposed. Many of these agree that data can have evidentiary value and that they can be used to support or to validate particular claims. The OECD report on *Principles and Guidelines for Access to Research Data from Public Funding* defines data as "factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings".[271] The data policy of the University of Edinburgh stresses similarly that data, "unlike other types of information, is collected, observed, or created, for purposes of analysis to produce original research results".[272] Research data are often considered to be a key element in the "chain of evidence" that underlies scholarly research in all disciplines.[273] As data commonly serve as the building blocks which ultimately enable scholars to construct an argumentation, they are generally viewed as semi-manufactures, rather than as the final products of a research project.

The definitions that have been cited can help to clarify the rationale of research data, but they do not offer a precise description of their nature. Definitions which focus more closely on their essence often emphasise that data are primarily descriptions or representations of the objects in a particular domain. Dervos and Coleman, for instance, discuss a distinction between 'facts' and 'data'. They explain that the term "fact" refers to "things done, that is, deeds or acts made into

[270] Deploring the confusion which frequently surround the term "research data" within the context of the humanities, Borgman urged digital humanists to provide better descriptions of the ways in which data are created, evaluated and used. See Christine L. Borgman, "The Digital Future Is Now: A Call to Action for the Humanities", s. 23.

[271] *OECD Principles and Guidelines for Access to Research Data from Public Funding*.

[272] "Research Data Management Guidance", <http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library/research-data-mgmt> (22 June 2014).

[273] Marlo Welshons, *Our Cultural Commonwealth: The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences*, (2006), p. 18.

something known" and to phenomena or events which "pertain to objective reality".[274] Data, on the other hand, "represent real world facts". They comprise "the outcome of measurements conducted in relation with real world phenomena".[275] A kindred view is offered by Marcia Bates. In an attempt to define data from a biological and evolutional perspective, Bates describes data as "that portion of the entire information environment available to a sensing organism that is taken in, or processed, by that organism".[276]

In the academic field of information science, many theorists view data as one layer in a larger hierarchy which connects the term to the related concepts of "information", "knowledge" and "wisdom". The series of models which represent the associations between these four concepts have been referred to varyingly as the "DIKW Pyramid" or the "DIKW Hierarchy". The model has undergone a number of revisions and extensions since its initial conception in the late 1980s,[277] but it is generally maintained that each layer derives from an underlying layer, and that the base of the hierarchy is ultimately formed by data. In most interpretations of the model, data are conceptualised as factual and non-interpreted representations of specific external phenomena. Russell Ackoff describes data as "symbols that represent properties of objects, events and their environments",[278] and Lieuw views data as "the storage of intrinsic meaning, a mere representation" whose main purpose is "to record activities or situations, to attempt to capture the true picture or real event".[279] In agreement with the etymology of the word "data", the theories and definitions that have been cited claim that data can be acquired by observing objects or events within an environment that is largely "given" a priori.

Definitions which depict data as representations of existing phenomena can help to embed the term within an existing discourse about the nature of the digital humanities. Many scholars view digital humanities as a field that is centrally concerned with the development of and the critical engagement with models. In his monograph *Humanities Computing*, Willard McCarty argues that the computer's value to humanities research stems primarily from the capacity to support "the

[274] This definition is quoted by the authors from Lawrence McCrank, *Historical Information Science* (Medford: Information Today, 2002), p. 627.

[275] Dimitris A. Dervos & Anita Sundaram Coleman, "A Common Sense Approach to Defining Data, Information, and Metadata", in Gerhard Budin, Christian Swertz, & Konstantin Mitgutsch (eds.)*Proceedings of the Ninth International ISKO Conference*, (Würzburg: Ergon-Verlag, 2006), pp. 51–58, p. 55.

[276] Marcia J. Bates, "Information and Knowledge: An Evolutionary Framework for Information Science", in: *Information Research*, 10:4 (2005), n.pag.

[277] For a good discussion, see J. Rowley, "The Wisdom Hierarchy: Representations of the DIKW Hierarchy", in: *Journal of Information Science*, 33:2 (February 2007).

[278] R. L. Ackoff, "From Data to Wisdom", in: *Journal of Applies Systems Analysis*, 16 (1989), p. 3.

[279] Anthony Liew, "Understanding Data, Information, Knowledge And Their Inter-Relationshipso Title", in: *Journal of Knowledge Management Practice*, 8:2 (2007), n.pag.

heuristic process of constructing and manipulating models".[280] McCarty defines a model as "a representation of something for purposes of study, or a design for realizing something new".[281] John Unsworth stresses likewise that computer-based humanistic research is fundamentally "a practice of representation" or "a form of modelling". In this form of research, "the computer is used as a tool for modelling humanities data and our understanding of it".[282] Humanities scholars are mostly interested in the nature, the historical development or the reception of cultural or artefacts, but these original artefacts are often in a format which complicates or even precludes a systematic digital analysis. When scholars aim to investigate texts which originated as physical objects, the relevant properties of these objects obviously need to be represented via bits before these can be studied computationally. Berry explains that the digital humanities "try to take account of the plasticity of digital forms and the way in which they point toward a new way of working with representation and mediation, what might be called the digital 'folding' of reality".[283] The idea that the concept of representation is crucial to the definition of data is also underscored in the Digital Humanities Manifesto 2.0, which was authored by Pressner and Schnapp. The text characterises the digital humanities as an area of research which "values the copy more highly than the original". Interestingly, the authors exploit the etymology of the word "copy", whose original meaning of "abundance" survives in the word "copiousness", to stress the extensive and widespread availability of digital surrogates of cultural artefacts.[284]

Computer-based literary research engages with text in a manner that is distinctly circuitous. Works of literature clearly form the ultimate objects of research, but critical analyses focus primarily on digital surrogates of these works. The DIKW model sets forth the view that data consist of surrogates of objects or of events, and this expedites the application of the term within the context of humanities research. Other characteristics of the DIKW model also complicate its pertinence to humanistic research, however. In a critique of the DIKW theory, Martin Frické writes that the model has a bias towards the natural sciences, and that it operates exclusively within the confines of positivism or empiricism. The model is based on the incorrect assumption that knowledge can only be obtained through a systematic analysis of sensory or observational data. The DIKW hierarchy allows no room for "unobservable ('theoretical') entities and properties"

[280] Willard McCarty, *Humanities Computing* (Basingstoke; New York: Palgrave Macmillan 2005), p. 23.

[281] Ibid., p. 24.

[282] John Unsworth, "What Is Humanities Computing and What Is Not?", in: Melissa Terras, Julianne Nyhan, & Edward Vanhoutte (eds.), *Defining Digital Humanities: A Reader*, 2013, pp. 36–37.

[283] David M. Berry, "Introduction", p. 2.

[284] Jeffrey Schnapp, Peter Lunenfeld & Todd Pressner, *The Digital Humanities Manifesto 2.0*, p. 14. The original text uses upper case for some of the central nouns, but the capitalisation has been removed in this quotation.

nor for objects or phenomena in the physical domain "for which no instruments of measurement exist".[285] Humanities scholars do not necessarily base their argumentation on factual or objective properties of observable phenomena, and the observations that are made by scholars are frequently of an interpretative or of a speculative nature. For this reason, conceptualisations of humanistic data can only be cogent when they acknowledge that data can be either factual or interpretative representations of physical or born-digital cultural artefacts.

Kitchin discusses a useful difference between "captured data" and "derived data". The first term refers to raw and unprocessed data which are generated "through some form of measurement such as observation, surveys, lab and field experiments, record keeping [...], cameras, scanners and sensors".[286] Captured data record neutral or unprocessed facts about an observable reality. Unprocessed datasets are often in a format that cannot be queried systematically, and, if this is the case, researchers need to restructure, classify or normalise the data.[287] Derived data, in contrast, are "produced through additional processing or analysis of captured data". In general, the level of neutrality or objectivity decreases with each phase of further processing.[288]

Building on a conceptual description provided by Davis et al., Unsworth identifies a number of generic characteristics of data. For the current discussion, three properties are of particular relevance. Data, firstly, are created to enable or to expedite particular types of analyses. They imply a "fundamental conception of intelligent inference",[289] and they sanction or recommend specific types of computational manipulations. Secondly, as is also stressed by McCarty, digital representations are typically based on an ontology. In philosophy, the term

[285] Martin Frické, "The Knowledge Pyramid: A Critique of the DIKW Hierarchy", in: *Journal of Information Science*, 35:2 (21 November 2008), p. 4.

[286] Rob Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences* (2014), p. 7.

[287] Liz Lyon, *Dealing with Data: Roles, Rights, Responsibilities and Relationships*, (Bath: 2007), p. 15.

[288] Research data can be categorised in many other ways. Kitichin discusses a distinction between primary, secondary and tertiary data. Primary data are "generated by a researcher and their instruments within a research design of their making". Secondary data, on the contrary, are "data made available to others to reuse and analyse that are generated by someone else". Tertiary data, finally, result from calculations and other forms of processing. Examples include "counts, categories, and statistical results". A similar classification is provided by Norman Blaikie, who argues that the distinction between primary, secondary and tertiary data also arranges the data according to the "notion of distance from the phenomena". Such distance is minimal with primary data, as they are "the result of direct contact between the researcher and the source". Tjalsma and Rombouts use the terms "primary data" and "secondary data" to distinguish between "empirical, observed, measured data" on the one hand, and "data derived from sources created previously". This description differs, nevertheless, from the definitions which were given by Kitchin and by Blaikie. Because of these differences he distinction between primary, secondary and tertiary data will not be used in this thesis.

[289] John Unsworth, "What Is Humanities Computing and What Is Not?", pp. 40–41.

"ontology" is generally used to refer to the "science or study of being",[290] but, in the context of humanistic modelling, the term is used more restrictively to describe the aspects of the original which are reproduced in the digital surrogate. It is an abstract theoretical conceptualisation of the original object to be modelled, enumerating the various qualities that need to be represented. A representation can never be fully representative, and it is inevitably based on a prior identification of the characteristics that are considered essential. As such decisions simultaneously imply a statement that the remaining aspects may be ignored, models "inevitably lie, by omission at least".[291] A third characteristic of digital surrogates is that they make use of a formal representation language. This language may also be referred to as a data format. Formatting is a "mechanism for describing data, i.e., for mapping concepts of a data model to digital objects such as files or memory".[292] A data format ensures that the aspects and the concepts which are considered relevant can be captured and manipulated on a digital device. The formal representation language consists of a range of symbols that can be used to express particular concepts, and a syntax that prescribes the manner in which these symbols may be combined into valid statements or data structures. Examples of data formats include plain texts, TEI-encoded texts, images, RDF-based annotations and formats in relational databases.[293]

This chapter provides a classification of the different types of research data that can be created and reused within literary informatics research. This classification is based on a consideration of the ontologies that underlie the various data formats. Ontologies and data formats are strongly linked, nevertheless. Because a particular data format can never represent all textual aspects, the choice of a format also implies specific ontological commitments. The nature of the ontology that underpins a particular surrogate can be investigated by comparing the properties of the surrogate to the qualities of the original. As a preamble to the characterisation of these data formats, the following section provides a detailed and generic discussion of the various aspects of the literary works that can be studied. The generic and expansive ontology that is presented in the next section can be used to analyse the more limited ontologies that are maintained within individual data formats.

Following a definition provided by Borgman, the original works that are studied can be referred to as sources. Sources, more specifically, comprise all the relevant materials that were created outside of the context of the research process, and which form the input for scholarly enquiry. Sources need to be distinguished from resources, which are the "data, documents, collections, or services that meet

[290] "Ontology", in *Oxford English Dictionary*, <www.oed.com> (16 May 2015)

[291] R. Davis, H. Shrobe & P. Szolovits, "What Is a Knowledge Representation?", in: *AI Magazine*, 14:1 (1993).

[292] Robert E. McGrath, *XML and Scientific File Formats*, (Urbana-Champaign: 2003), p. 6.

[293] These formats are discussed in more detail in the following sections.

some data or information need".[294] Resources are created by scholars themselves in the course of their research. The nature of research data can understood by considering their relation to the sources they mimic.

4.3. Sources

Some of the generic properties of textual sources can be analysed effectively by making use of the conceptual entity relationship model which was developed by the *International Federation of Library Associations* (IFLA), under the name *Functional Requirements for Bibliographical Records* (FRBR).[295] The FRBR model, which was devised originally to represent the various bibliographic levels which may be present in a library catalogue,[296] distinguishes four cardinal terms. A "work", firstly, is "a distinct intellectual or artistic creation". It can be understood as a Platonic representation of a particular creation, because "there is no single material object one can point to as the work". The second entity in the FRBR model is the "expression". It is an "intellectual or artistic realization of a work in the form of alpha-numeric, musical, or choreographic notation, sound, image, object movement, etc., or any combination of such forms". The original text as written by the author is considered to be an expression. The sequence of characters that is produced is an embodiment of a particular work. Subsequent new editions or translations of this text establish a new sequence of characters, and they are consequently new expressions of this work. A 'manifestation', thirdly, is an expression which has been presented on a particular medium, using a specific typography.[297] When a single edited version is made available multiple times with a different typographical appearance or on different media, these are all considered to be different manifestations. An 'item', finally, is "a single exemplar of a manifestation". The 'item' is the only concrete class of objects in the FRBR model, since the three additional levels are abstract concepts which can be used to make statements about the ways in which items can be connected. Digital surrogates of printed works are necessarily representations of a particular item. Such an item contains a particular string of characters, and it is presented through a particular typography.

Whereas the FRBR model can be applied effectively to clarify the differences between separate printed editions, the conceptualisation may be less appropriate for literary texts that are transmitted via other media. The distinction between expression, manuscripts and item is problematic, for instance, for literary texts

[294] Christine Borgman, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*, p. 122.

[295] *Functional Requirements for Bibliographic Records: Final Report*, (2009).

[296] Karen Coyle, "FRBR, the Domain Model", in: *Library Technology Reports*, 2010, p. 21.

[297] Views vary on whether or not a change in medium actually causes the creation of a new manifestation.

which were created before the invention of moveable type, and which have survived as handwritten sources. When scribes copied a manuscript, they inevitably introduced alterations in the text, thus establishing a new expression. As each copy was produced in the hand of an individual scribe,[298] each new manuscript may also be seen as a separate manifestation. Thirdly, each manuscript is also unique, and this renders the concept of the item, as a level distinct from the manifestation, inconsequential. In the case of a manuscript, only two levels can reasonably be distinguished, since the concepts of expression, manifestation and item coalesce. The distinction between manifestations and items may also be disputed in the case of electronic sources. Van der Weel notes that the essential virtuality of digital files admits the possibility to produce an unlimited number of copies which are indistinguishable from the original. The copy and the original are so much alike, in fact, that it may no longer be appropriate to refer to the duplicated file as a copy.[299] The concept of the copy appears to be salvaged, however, by recent technical developments. A growing number of applications enable readers to annotate and to manipulate specific copies of digital resources.[300] The personal annotations which were added by a particular reader, and which are clearly unique to a particular file, can be relevant from a scholarly perspective. In the case of such annotated digital files, reinstating the item as a separate entity can be justified.

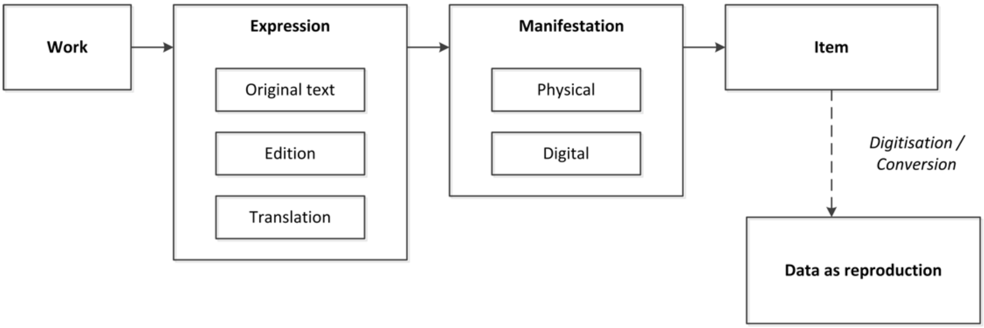


Image 4.1. Entities of the FRBR model

[298] In scriptoria which had implemented the pecia system, a single text can also be written in the hands of multiple scribes.

[299] Adriaan van der Weel, *Changing Our Textual Minds : Towards a Digital Order of Knowledge*, p. 150.

[300] Applications such as iAnnotate, Browzine and Mendeley form important examples.

Whereas the different levels that are proposed by FRBR cannot be applied equally suitably to all the sources that can be investigated by literary scholars, the model's main value for the current discussion lies in the fact that it identifies a number of crucial analytic components of the concrete physical or digital object whose properties are mimicked in a surrogate. A first important observation is that a textual source, as an expression of a work, comprises a particular sequence of characters. Peter Shillingsburg refers to these characters as the "lexical codes", and explains that these consist of the "the more usually acknowledged aspects of a text, letters, accents, and punctuation".[301]

FRBR also draws attention to the fact that the original artistic expression can be extant in different editions. Literary research can incidentally take place by inspecting the original manuscripts of an author, but the majority of critics base their findings on published editions of literary works. While, in the natural sciences, research typically concentrates on phenomena which are given and which exist independently of the observer, literary scholars can intervene directly in the domain they study by actively shaping the nature of textual sources. The aim of editing, broadly speaking, is to make texts more accessible or more useful to specific audiences by organising, correcting or even paraphrasing original works. Scholarly editing is a specific type of editing in which the various modifications that are made to the primary sources are based on scholarly research. The type of research that underpins such scholarly editions is commonly referred to as textual criticism.[302] The alterations that are introduced mostly follow the rules and the standards that are prevalent within specific editorial frameworks or traditions.[303] Elena Pierazzo explains that the different approaches that are followed in scholarly editing are defined, amongst other aspects, by "the way they handle the evidence offered by primary sources", by "the way they reconcile contrasting readings from different sources" and by "the importance given to authorial intention".[304]

Tim Machan discusses a useful distinction between lower criticism and higher criticism. The former type of criticism mostly entails "the establishment of literary, social, and cultural contexts and thus subsumes biography, bibliography, and

[301] Peter Shillingsburg, *From Gutenberg to Google: Electronic Representations of Literary Texts*, p. 16.

[302] Kathryn Sutherland, "Being Critical: Paper-Based Editing and the Digital Environment", in: Marilyn Deegan & Kathryn Sutherland (eds.), *Text Editing, Print and the Digital World*, Farnham: Ashgate, p. 13.

[303] Examples of such frameworks include stemmatics, which is frequently the norm for editions of medieval manuscripts, Walter Greg's copy-text theory, which has often been applied to early modern print materials, and genetic criticism, which aims to represent the writing process that was followed by an author. See Elena Pierazzo, *Digital Scholarly Editing: Theories, Models and Methods* (Farnham: Ashgate 2015), p. 11.

[304] Ibid.

textual criticism”. [305] Higher criticism, by contrast, is directed more specifically towards explanation and literary interpretation. Lower Criticism is “commonly viewed as the more factual or ‘scientific’” approach as it establishes “numerical, analytical, and categorical information which is used to define historical realities”.[306] Machan adds that there is typically a reciprocal relation between textual criticism and interpretative research. The results of the exertions of editors and textual critics form the sources of the literary critic. Textual criticism aims to produce the semi-manufactures on which literary critics can base their arguments, and the lower form of criticism serves as “a stepping stone to the Higher, and sometimes more important, one”.[307] Conversely, the needs and the aims of literary critics also provide a context for the activities in the field of lower criticism. Particular interpretations of a text may affect the way in which textual critics approach later editions of this text.

Historical-critical editions mostly aim to represent the author’s original intentions as directly as possible, and to avoid any bias on the part of the editor. Many of the core tenets of the historical-critical approach were originally formulated by the philologist Karl Lachman, who proposed a stemmatic approach aimed at tracing the lineage between the various witnesses of a text. The objective of the approach, ultimately, is to expose the work in its original form, free from errors and omissions.[308] In a similar vein, Walter Greg proposed that critical bibliography should serve as an anchor for the subjective activities of the literary critic, and stressed that editors should refrain as much as possible from making critical interpretations.[309] More recent theorists of the field of textual criticism have stated that such neutrality is unfeasible, and have embraced the fact that editions invariably reflect the preferences of textual critics. The objective to reconstruct an original and authoritative version of a text has now largely been replaced by the conviction that all extant versions of the text have value in themselves. The New Bibliography movement, for instance, abandoned the notion of textual idealism, and recognised that editors generally write much of their own interpretations into a text. Paul Zumthor's influential “mouvance” theory suggested likewise that all edited texts should be placed in their social context, as each textual variant is part of the reception history of the text. Machan emphasises that “there can be no value-free textual criticism”, and Peter Shillingsburg agrees that “the compilation of a scholarly edition is the interpretive best thinking of an editor and is not the

[305] Tim William Machan, “Late Middle English Texts and the Higher and Lower Criticisms”, in: Tim William Machan (ed.), *Medieval Literature: Texts and Interpretation*, Binghamton: Center for Medieval and Early Renaissance Studies 1991, p. 4.

[306] Ibid.

[307] Ibid., pp. 4–6.

[308] Paul Maas, *Textual Criticism* (Oxford: Clarendon Press 1958), p. 3.

[309] G Tanselle, *Textual Criticism since Greg : A Chronicle, 1950-1985* (Charlottesville: University Press of Virginia 1987).

establishment of a text for all time".[310] Since literary works are typically available in multiple editions, and since each edition, to a higher or a lesser degree, entails a particular reading of a text, literary scholars who aim to study a work digitally need to be critical with respect to the edition they aim to model.

On paper, an edition generally publishes a single variant of the text, since a paper-based compilation of all variants would be impracticable and economically unviable. In the digital realm, the mouvance theory can manifests itself in practical terms in what Vanhoutte refers to as the "maximal edition".[311] The term refers to an edition in which the entire transmission history of a text is represented, and in which all available variants are included. In most cases, maximal editions can only be created in a digital form. To ensure that users can purposefully navigate a large numbers of variants, maximal editions must exploit the malleability of digital texts, in an interface in which fragments from selected witnesses can be shown side-by-side. According to Sutherland, however, the possibility to include each extant witness only has limited worth. If the editor omits the act of selection, the burden of having to select a specific version shifts to the end-user, who is normally not in a position to evaluate the significance of the various options. For this reason, Sutherland refers to such digital editions as "recyclable wastebanks".[312]

Within a particular expression of a work, a number of textual aspects can be identified. The lexical codes of the text, which are often particular to a given edition, contains a logical structure. The characters that make up a text are divided over distinct logical units such as chapters, sections, paragraphs and sentences. The various logical units are normally used in the service of a rhetorical structure. Units such as paragraphs and section often have a specific function in the narrative or in the overall the argumentation of the text. The rhetorical structure enables readers to trace, for instance, how certain conclusions follow from premises which have been introduced earlier. This logical structure has usually been conceived before the text was cast onto a particular medium. There is also a second set of structural units, however, which is brought into existence by the placement of the full text onto paper. Examples in the latter class include running titles, page numbers and title pages. These components were generally not conceived of by the author, and they have usually been added to facilitate the navigation through the text. Gerard Genette explains that these structural components from part of the paratext. Paratextual units which are visible within the publication, and which

[310] Peter Shillingsburg, *From Gutenberg to Google: Electronic Representations of Literary Texts*, p. 171.

[311] Edward Vanhoutte, "Every Reader His Own Bibliographer - an Absurdity?", in: Marylin Deegan & Kathryn Sutherland (eds.), *Text Editing, Print and the Digital World*, Farnham: Ashgate 2009, p. 111.

[312] Kathryn Sutherland, "Being Critical: Paper-Based Editing and the Digital Environment", p. 26.

result from the text's *mise-en-page*, are referred to, more specifically, as the peritext.[313]

The actual running text, not including the peritext, displays a linearity. In this context, the term linearity refers to the notion that the units of the language need to be processed in a fixed order, and that the units, to a large extent, derive much of their meaning from their placement within from this particular context. In his *Course on General Linguistics*, De Saussure has noted that such a linear progression is characteristic of all texts in natural language. He explains that the linguistic signal has a "temporal aspect, and hence certain temporal characteristics: (a) it occupies a certain temporal space, and (b) this space is measured in just one dimension: it is a line".[314] Linearity is characteristic of both written and spoken texts. In the case of a paper-based text, "a spatial line of graphic signs is substituted for a succession of sounds in time".[315] For readers, it is generally difficult to locate units of information instantaneously and separately from their context. To fully understand the meaning of a narrative, readings generally need to consider the broader context.

Next to the fact that a text consists of specific string of lexical codes, the FRBR model also recognises that textual sources have a typography. The function of typography is generally to clarify the logical structure, which includes the components that belong to the peritext. Items such as paragraphs, block quotes, section heading and footnotes are normally rendered distinct visually, and this enables the reader to decipher the logical category of each segment. The nature of the units that belong to the peritext is often clarified via the positioning of these units on the page. Running titles, for instance, can be recognised as such because they are placed at the top of the page.

In FRBR, typography is not considered to be part of the original artistic expression. The definition of the term expression excludes "aspects of physical form, such as typeface and page layout, that are not integral to the intellectual or artistic realization of the work as such". According to the model, the work of the typographer may be compared to that of a conductor of a piece of music, who essentially provides an interpretation of a work that was conceived originally by a composer. If the same content is published multiple times, but with a distinct typography, these two texts still belong to the same expression. A related argument can be found in the article "What is a Text, Really?", in which DeRose et al. consider "the question of essentials: What is it which, if changed, makes a document essentially different, and what is it which can change, yet a document remains 'the same?'". The authors conclude that the words of the text constitute the

[313] Gerard Genette & Marie Maclean, "Introduction to the Paratext", in: *New Literary History*, 22:2 (2010), pp. 263–264.

[314] Ferdinand de Saussure, *Course in General Linguistics* (London: Duckworth 1983), p. 69.

[315] Ibid., pp. 69–70.

genuinely “meaningful units”. When the words change, a new text originates, while “adjustments of typography” appear to be “superficial and transient rather than essential”.[316]

Various authors have argued, nevertheless, that typography makes a crucial contribution to the overall experience and significance of text. Don McKenzie, for instance, explained that “the material form of books, the non-verbal elements of the typographic notations within them, the very disposition of space itself, have an expressive force in conveying meaning”. The strong connection between form and content is underscored by the etymology of the word ‘text’, which indicates “a process of material construction”.[317] In a printed text, content and typography are woven together into a single indelible unit, and the form actively contributes to the production of the text’s message. According to McKenzie, texts ought to be studied as “recorded forms” which have originated at a particular location and at a particular time. In printed books, specific aspects of the text’s layout are indicative of a particular reading and, in turn, these influence the manner in which new readers interpret the text.

The position that typography is transparent and semantically void is confounded severely by literary texts in which the author has explicitly used the textual form as an expressive element. The typography of a text strongly affects the manner in which readers experience this text. Lennard stresses that decisions about the layout of a text are not made exclusively by printers or by publishers, since authors increasingly begin to employ this aspect in order to convey meaning.[318] Works in the literary genre that is commonly referred to as “concrete poetry”, for instance, have been “composed with specific attention to graphic features such as typography, layout, shape, or distribution on the page”.[319] Eugen Gomringer’s “Silencio” and Decio Pignatari’s “Bebe coca cola”[320] exhibit the notion of isomorphism, which is a genre of poetry in which shape and meaning are considered identical. Other examples of texts in which the typography has a clear semantic function can be found in the works of George Herbert, Paul van Ostaijen en Dom Silvester Houedart. In Herbert’s pattern poems, for instance, of which “The Altar” and “Easter Wings” are probably most widely known, the shape that is formed by the words on the page ingeniously depict and support the subject matter. In digital surrogates of such concrete or visual poems, it is essential to

[316] Steven J. DeRose et al., “What Is Text, Really?”, in: *Journal of Computing in Higher Education*, 1:2 (1990).

[317] Donald Francis McKenzie, *Bibliography and the Sociology of Texts* (Cambridge: Cambridge University Press 1999), p. 3.

[318] John Lennard, *The Poetry Handbook: A Guide to Reading Poetry for Pleasure and Practical Criticism*, p. 47.

[319] *The Princeton Encyclopedia of Poetry and Poetics*, p. 294.

[320] Both poems are discussed in *The Princeton Encyclopedia of Poetry and Poetics*.

ensure that the characteristic aspects of the form can be retained, since these texts would otherwise be bereaved of much of their expressive value.

McKenzie's "recorded forms" consist of lexical codes combined with a particular typography, and, accordingly, they correspond to 'manifestations' in the FRBR model. In the context of the printed book, the expression and the manifestation inevitably coincide in a physical item, as it is not physically possible to separate lexical codes from the typography. Terms which, in the context of physical publications, largely refer to abstract concepts may actually be used to describe distinct types of files in the context of the digital medium. A plain text which consists exclusively of Unicode characters can be considered an expression of a text. Such an expression can be marked up typographically, using, for instance, a DTP program, and the result will then be a manifestation of this text. Similarly, when a text is marked up using XML, a stylesheet can be applied to this document in order to render the encoded content in a specific typographical form. Because of this separation of content and form, readers of digital sources can be confronted with texts in which the typography is intrinsically unstable. A distinction can be made between fixed formats, such as PDF or TIFF, and rescalable formats, such as HTML or EPUB. When texts are disseminated in rescalable formats, layout artists only have limited possibilities to fix the presentation. In the case of HTML, for instance, the typographical appearance often depends for a large part on the particular settings of the browser in which users read a text. In eBooks based on EPUB, users are often able to personalise the presentation of the text, and to change font faces, letter sizes or background colours. As readers can flexibly weave and re-weave the contents into new forms, rescalable eBooks complicate MacKenzie's notion of texts as recorded forms. The plasticity of digital sources undermines the typography's ability to convey and to support a particular reading.

The current enumeration of the cardinal properties of textual sources concludes with a brief mention of a number of obvious characteristics, which need to be mentioned for the sake of completeness. A text, evidently, has semantic contents. The text invariably addresses specific topics, and it frequently references items in specific generic categories of information, such as personal names, geographic terms or book titles. As discussed in the previous chapter, literary texts can also make use of literary devices, such as metre, rhyme, alliteration and imagery. Texts, furthermore, have linguistic properties. Literary writing often employs a register of language that differs in a number of important ways from common colloquial language. Literary analyses may, for this reason, also focus on occurrences of syntactical categories or of particular grammatical constructions. Words, when read aloud, are also associated with sounds. Literary devices are often based on a skilful use of such sounds. Literary research, and, in particular, analyses of poetry, often concentrate on the phonetic aspects of a text. Words in the English language are generally used in conjugations and in declensions. In some cases, it can be useful to replace inflected forms with lemmas, which are their base dictionary forms.

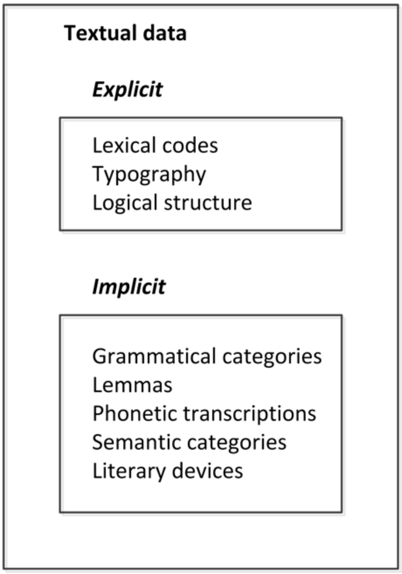


Figure 4.2. Core aspects of texts

The properties that have been discussed in this section, and which may potentially be mimicked in a surrogate, are summarised in figure 4.2. Literary research focuses on texts written by literary authors. Importantly, the verbal construction that is studied is rarely a fully neutral representation of the words that were written by the author, as the nature of the lexical codes is often informed by the editor's opinion about the nature and the purpose of the literary text. Additionally, while the text primarily produces its meaning via the lexical codes, the form of the text frequently contributes strongly to the overall effect of the text. This is particularly the case for concrete or visual poetry. Texts often bear the marks of subjective interpretations performed by textual critics or by typographers. The published edition, nevertheless, is an objective artefact which can serve as the basis for subsequent digital scholarship.

4.4. Captured data

Computer-based literary research typically commences with a process in which a collection of sources which were intended to be read by human beings are transformed into artefacts that can be studied computationally. The fact that digital scholarship demands a conversion is obvious in the case of texts which were initially cast onto an analogue medium. Converting physical sources into digital objects inexorably discards aspects that are tied uniquely to the physical form, such as the paper quality, the dimensions, the tactile sensation and the smell of the object. Such characteristics can be studied exclusively by consulting the original. A conversion from analogue into digital is obviously not needed for born-digital texts. In most cases, nevertheless, sources which are natively digital still need to be converted, as such sources often contain navigational or typographical aspects which may hinder a systematic analysis. The form that was devised for human readers is not necessarily suitable for scholarly purposes.

This chapter makes a broad distinction between data formats that may be used most appropriately for the representation of observable and explicit aspects of texts on the one hand, and formats that are more suitable for the description of aspects which are implicit or latent in the original works on the other. Data which describe observable aspects of the original object can be classified as “captured data”. In the previous section, eight cardinal aspects of texts have been identified, and two of these, the lexical codes and the typography, have an explicit and observable presence. About the text’s typography, it must be stressed that, whereas the formal features are manifest, the semantics associated with these presentational devices is implicit. The aim of the typography is mostly to clarify the text’s logical structure, but there are no explicit markers which unambiguously identify the nature of the various logical components. A group of characters separated by hard returns may constitute a paragraph, a stanza, an epigraph or a block quote, among other options. In most cases, a human reader can decode the typographical conventions flawlessly, in the same way as he does the lexical codes.[321] Since the nature and the function of these units are not declared explicitly, these cannot be identified directly by computers. Because of the implicit semantics of typography, it may be argued that the logical structure is partly manifest and partly latent.

The plain text format can be used to reproduce the lexical codes of a text. Within this format, the characters that are contained within the work can be accessed separately. In a sense, the plain text is a reversal of the process that was initiated by the invention of moveable type. While Gutenberg used a limited set of cast characters to produce fixed and stable objects, many of the affordances of the electronic text are based on the fact that its letters, digits and punctuation marks can be manipulated separately. As was discussed, machine-readable texts are made available by an increasing number of libraries or commercial organisations. The vast text base that has been assembled as part of *Project Gutenberg*, for instance, is a common source of primary data for textual scholars. About this particular resource, Peter Shillingsburg notes that it often has insufficient information to ascertain that the texts are sufficiently accurate. Additionally, it is frequently

[321] Adriaan Van der Weel explains that human readers “truly deserve to be called *homo typographicus*” because of the “astonishing ease with which we are capable of assessing unconsciously the purport of textual messages without even reading a word of the actual text”. See Adriaan van der Weel, *Changing Our Textual Minds : Towards a Digital Order of Knowledge*, p. 69.

unclear whether or not editors have chosen a source text that has a degree of authority or historical importance.[322] If no accurate and authoritative plain texts are available, scholars may produce plain texts themselves, by transcribing the text, or through the use of OCR software.

Captured data are rarely fully accurate and seldom entirely objective. While digitisation projects often strive to produce reliable and unadulterated representations of the original sources, it is usually impossible to avoid the introduction of alterations during the conversion process. The structural inadequacy of the results of OCR scanning, for instance, have been documented extensively.[323] The results can be particularly poor for books containing uncommon font types or for books with a low print quality. In the case of handwritten materials, the texts usually need to be transcribed manually. During such processes, subjectivity can never be avoided entirely. Human scholars may make typing errors unknowingly, but they may also change the text more consciously for the purpose of specific editorial interventions. When a particular hand is difficult to read, making a transcription obviously demands interpretation. In addition, medieval manuscripts often contain various abbreviated words and phrases which transcribers may choose to expand, but views may vary on what these abbreviations actually stand for. In general, a wide range of choices need to be made during the conversion of analogue sources into digital resources, and different persons may also take different decisions.

When texts that have originated on a paper medium are scanned and converted into machine-readable text, this is not a simple migration of content, as the digitised version differs from the original in a number of important ways. Research which is based solely on data in the plain text format crucially disregards the text's typography. As plain texts are essentially immaterial and formless resources, the use of this format is problematic for literary works in which the typographical presentation is an inherent part of the artistic expression. Plain machine readable text, furthermore, has a strict linearity. It fundamentally consists of one long concatenation of characters and spaces. In contrast to the physical page, plain texts cannot encode a difference between text and peritext via positioning. If peritextual aspects such as running titles, page numbers, and section headings are considered irrelevant, these need to be removed cautiously.

As was explained above, digital scholarly editions may be produced which can facilitate access to all extant witnesses of the text. It may be useful from a scholarly perspective to have a complete overview of the genesis of texts, but such a multiplicity of textual variants can complicate computer-based textual analyses. Many of the operations in the context of literary informatics are based on counts of

[322] Peter Shillingsburg, *From Gutenberg to Google: Electronic Representations of Literary Texts*, p. 21.

[323] See, for instance, Martin Volk, Lenz Furrer & Rico Sennrich, "Strategies for Reducting and Correcting OCR Errors", in: Caroline Sporleder, Kalliope Zervanou, & Antal van den Bosch (eds.), *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, Berlin: Springer 2011.

the occurrences of words, and if two different variants of one single text are both processed, this may obviously distort the statistics and overrepresent the frequencies of certain words. Scholars should therefore consider the texts that they include in their studies very carefully, and select the single variant which, in their view, best represents the text. Interestingly, it can be seen that, in this way, the use of algorithms promotes a return to a more idealistic view on textuality. The digital medium initially caused a decrease in the importance of an ideal and authentic version of the text, as limitations with respect to space in the printed editions did no longer apply.[324] While, in a digital edition, it becomes feasible to give expression to the view on editing as an open-ended discussion, the inevitable duplications that occur in such maximal editions clearly pose difficulties for scholars who aim to analyse the texts algorithmically.

Plain texts cannot reproduce aspects of typography such as font family, font size and colours, but they generally do include spaces, punctuation marks, capitals and line breaks. Such features of the text can be used, to some extent, to recognise the logical structure of the text. Whereas text mining has been described in Chapter 1 as a field which extracts information from unstructured data, it can be observed that texts generally do contain a degree of structure. Text mining may therefore be described more aptly as a field which exploits the formal and explicit clues which are present in the text to convert an implicit or inconsistent structure into an explicit machine-processable information structure. In most cases, text miming algorithms can only detect the presence of distinct logical units, nevertheless, and they cannot identify the nature or the function of these units. In most cases, including information about the function, the nature or the rhetorical effect of textual units demands a transition from data that represents explicit aspects to data that are descriptive and interpretative.

The lexical codes of a literary work can alternatively be represented via images of the printed page, in formats such as TIFF, JPG or PDF. A crucial disadvantage of these formats, however, is that the characters of the text cannot be manipulated until they are converted into machine readable text via OCR. Images can also reproduce the text's typography, which is a second textual aspect whose presence is manifest rather than latent. Images primarily provide a convenient access to digital versions of the original works, and they enable scholars to investigate the typographical aspects of the scanned pages on an individual basis. The development of algorithms for the automated analysis of the typography of large volumes of pages seems more challenging. Scholars who are interested in the

[324] Vanhoutte notes that the "electronic paradigm in scholarly editing has almost exclusively focused on the advantages of the size and economics of available storage capacity", and adds that the "digital archive as expanded text has in some cases jostled the one text away in favour of the multitude of many texts". Edward Vanhoutte, "Every Reader His Own Bibliographer - an Absurdity?", p. 109. In an important sense, the technical possibility to provide access to a large number of witnesses erodes the necessity of having to designate any single text as the most authoritative expression.

effects of typography generally concentrate on aspects such as indentations, font types and font sizes. At present, such characteristics cannot easily be extracted via digital image processing.[325]

4.5. Annotations

Computers can only process data which are present in an explicit form, or which can be derived consistently and unambiguously from other aspects which are explicit. Texts in natural languages, and works of artistic creation in particular, often have characteristics that complicate systematic querying. One important difficulty is that many of the characteristics that are of relevance to scholars are implicit.[326] An additional difficulty which hinders analysis is that natural language texts generally contain homonyms and synonyms. One concept may be referred to via distinct terms, and at the same time, one particular word may also refer to many different concepts. As the computer demands "complete explicitness and absolute consistency",[327] the impetus to study cultural objects via the digital medium implies the necessity to create digital surrogates in which all the properties that are implicit and imprecise in the original have been given an explicit and unambiguous expression. Smith explains similarly that, while the focus of the computer is on the literal characters that constitute the texts by default, the width of analytic procedures can be expanded if scholars encode "physical as well as semantic characteristics ... into symbol sequences parallel to the textual sequence". Smith suggests that such supplementary categories may be viewed as "strata that are parallel to and 'above' the textual sequence".[328]

[325] Lev Manovich has developed a method for the automated analysis of the style used in comic books by extracting data about "contrast, ... of texture and fine details, number of lines and their curvature". (Lev Manovich, "How to Compare One Million Images?", in: *Understanding Digital Humanities*, New York: Palgrave Macmillan 2012, p. 262.). These aspects do not seem relevant for the study of developments in typographical design, however.

[326] Among other aspects, plain text lack explicit data about the logical structure of the text. Because of this absence, it is impossible to distinguish the characters that appear in running titles or in the title pages from the actual body text of the literary work. In an acrimonious critique of the digital humanities, Adam Kirsch has argued that this inability to distinguish text from peritext has resulted in clear cases of unsound reasoning in studies based on big data. He illustrates his claim using a study conducted by Erez Aiden and Jean-Baptiste Michel. On the basis of the observation that references to specific years are most common in books published in that same year, the researchers claim that there is a general decline in historical awareness. Kirsch notes that this finding can be explained through the simple fact that the convention to print the year of publication on the copyright pages of books became more and more common. See Adam Kirsch, "Technology Is Taking Over English Departments: The False Promise of the Digital Humanities", in: *New Republic*, :May 2 (2014), n.pag.

[327] Willard McCarty, "Modeling: A Study in Word and Meaning", in: Susan Schreibman, Ray Siemens, & John Unsworth (eds.), *A Companion to Digital Humanities*, Blackwell, p. 258.

[328] John B. Smith, "Computer Criticism", 19–20.

Since the original sources do not include unequivocal markers of all the aspects that may be studied, such labels, if they are needed, must be supplied by the scholar. Implicit aspects can be made explicit via a process which, in this thesis, is referred to generically as "annotation". Structured annotations can be recorded via XML-based encoding. Stephen Ramsay notes about XML encoding that it can serve as "an elaborate scaffolding by which the vagaries of continuity can be flattened and consistently recorded".[329] The textual fragments to be annotated must initially be identified. Annotations may target words, phrases, sentences or texts in their entirety. XML offers a mechanism whereby fragments can be identified by placing a start tag and an end tag around the fragment to be annotated. The delineated fragments can subsequently be described by supplying a descriptive term either as the name of an XML element, or as the value of an XML attribute. Such a descriptive value can characterise the fragment as an instance of a particular phenomenon.

Within the humanities, the most widely used XML-based encoding language is the Text Encoding Initiative (TEI).[330] The standard was developed by a consortium of scholars in the humanities and the social sciences, and it currently consists of more than 500 descriptive terms.[331] Using the TEI, scholars can explicitly describe structural and semantic components of texts, such as paragraphs, sentences, place names, personal names and book titles. The TEI is a flexible and modular standard and was developed to support various forms of textual research. The standard provides facilities for "texts in any natural language, of any date, in any literary genre or text type, without restriction on form or content". [332] The development of the TEI represents a major effort to standardise descriptive practices in the context of textual scholarship.

The TEI is an instance of an embedded mark up technique. The descriptive codes are interspersed with the lexical codes of the texts, and annotations can be distinguished from the text itself as a result of specific notational conventions. Annotations can alternatively be captured using data formats in which the descriptive values are separated from the plain text. Such external annotations can be recorded, for instance, using an entity-relational model that is implemented in a relational database. In this particular class of data formats, the descriptive values are contained in the cells of the various tables. The column names provide explicit information about the properties that are being described. External annotations can also be stored as statements based on the Resource Description Framework

[329] Stephen Ramsey, "Algorithmic Criticism", p. 8.

[330] James Cummings, "The Text Encoding Initiative and the Study of Literature", in: *Blackwell Companion to Digital Literary Studies*, Oxford: Blackwell 2007.

[331] An XPath query of the tei_all.xsd TEI schema which counts the use of <xs:element> returns a total number of 517 occurrences.

[332] "P5: Guidelines for Electronic Text Encoding and Interchange", <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/AB.html> (27 January 2014).

(RDF). RDF is a technique that can be used to formulate generic statements. It does not provide a vocabulary for such statements in itself, but it offers a general framework or a data model for making statements, which centrally consists of a structure containing a 'subject', a 'predicate' and an 'object'. According to Lee, this tripartite structure can be used to "say anything about anything".[333] The Open Annotation Collaboration (OAC) and the conceptualisation that was developed for the Nanopublications schema are two RDF-based techniques for the creation of annotations.[334] In the case of RDF-based statements, the object of the statement generally contains a descriptive value for particular text fragments, and the predicate often clarifies the aspects which are being described.

Embedded mark up schemes such as TEI can delineate text fragments via the placement of opening and closing tags, but systems for the recording of external annotations often lack a standardised method for the identification of the target of the annotation. Specific annotations about, for instance, the date of creation or about the overall theme apply to the work as a whole, and, in such cases, a reference may be included to the text in its entirety. When a finer level of granularity is required, however, data formats for the creation of external annotations can be combined with standards for the addition of embedded mark up. In the TEI, the passages that are outlined via the mark up can be identified via an @id attribute, and those passages can consequently be referenced in records in relational databases or in RDF-based statements.

Data formats for the storage of annotation can be used to describe aspects of the text which are latent, such as syntactic categories, phonetic transcriptions, lemmatised forms and literary devices. XML-based annotations can also be used to describe aspects which are wholly or partly explicit in the text, such as the typography or the logical structure. Mark up which concentrates on the typographical aspects of texts is referred to by Coombs et al. as "procedural mark up".[335] Such procedural mark up has been used in natively digital sources encoded in HTML or in EPUB. Mark up tags can alternatively identify particular units such as verse lines, paragraphs, titles or block quotes, without specifying the typographical appearance of these units. Next to the intrinsic characteristics that were enumerated in section 4.3, annotations may also capture extraneous or contextual data about the texts. For studies in the field of literary history, it is often necessary to collect data on the creation or on the publication of the text. Aspects such as the author, the title, the imprint and the general subjects can often be retrieved from library catalogues. Such supplementary data about the creation and the publication of texts can be referred to as metadata or as 'bibliographical data'. Additionally, in

[333] Lee Lacy, *OWL: Representing Information Using the Web Ontology Language* (Victoria BC Canada: Trafford 2005), p. 75.

[334] These two techniques will be discussed in more detail in Chapter 6.

[335] James H. Coombs, Allen H. Renear & Steven J. DeRose, "Markup Systems and the Future of Scholarly Text Processing", in: *Communications of the ACM*, 30:11 (1987).

the case of online sources, it is also possible to collect data about the usage of particular texts. The server software that manages the access to a site normally keeps track of various aspects of the clients that request content, such as their geographic location, and the search terms that were used to find the text. Such data about the usage of documents can form a valuable source of information for research on the reception of literary works.

Structured annotations of the type that has been discussed can capture data which are supplementary, in the sense that they are not present in a manifest way within the original source. As the transformation of implicit aspects into explicit data almost inherently demands interpretation, the annotations that are furnished by individual scholars frequently have a subjective character. Different scholars may produce different observations about identical fragments. Johanna Drucker stresses that humanistic data are critically co-dependent on the observer, and that they are subsequently marked by idiosyncrasy and by ambiguity. Because of the constructivist nature of observations in the humanities, Drucker also objects to the term data, whose etymology suggests the existence of phenomena which are given a priori. Rather than viewing data as objective and pre-existing, they should be conceptualised as *capta*, or as observations which are taken.[336] Terms such as *facta* or *constructa* would perhaps be even more apt to emphasise the researcher-specific and the inventive nature of structured annotations.

It must be stressed, nevertheless, that this characterisation of data as subjective constructions cannot be applied to all the data within the field of literary informatics. In this chapter, a distinction was made between data which model observable properties of primary sources on the one hand and data which convert implicit and ambiguous aspects into structured and explicit annotations on the other. The aim of the former type of data is generally to objectively represent aspects of bibliographic items, such as printed editions or original manuscripts. A particular edition, once published, becomes a source which literary scholars can treat as a given. Plain texts and images may reasonably be viewed as objective data rather than as observer-dependent data. Additionally, while annotations that describe implicit aspects are often interpretative in nature, scholars may frequently arrive at a degree of intersubjectivity with respect to the aspects that are party or wholly explicit, such as the typography or of the logical structure. In the case of digitised texts, the correctness of the encoding that focuses on the logical structure of the text can be verified, to some extent, by inspecting occurrences of titles, paragraphs, line breaks or verse lines in the original. This verifiability confers a degree of objectivity on this type of encoding. Descriptive encoding which concentrates on the logical structure can never be fully objective or uncontested,

[336] Johanna Drucker, “Humanities Approaches to Graphical Display”, in: *Digital Humanities Quarterly*, 005:1 (2011), par. 3.

however, as human interpretation will still be needed to decode typographical devices into the structural components they clarify.

Annotations can be created in two ways. Human scholars can firstly choose to make some of their understanding of texts available manually. When critics provide data by hand on a case-by-case basis, this means that the heuristics for the recognition of specific features do not have to be programmed into software applications. The manual creation of secondary data results in what Christoph Schöch refers to as "smart data". Schöch argues that the manual creation of digital data does not differ dramatically from traditional work in the humanities, as it mostly demands a meticulous and labour-intensive close reading of texts. Because of the dependence on manual work, smart data "does not scale well".[337] Alternatively, annotations can also be produced by making use of text mining. More concretely, this entails the application of algorithms which have been designed to detect specific items of interest within the text. The output of these algorithms can be captured as annotations.

Whereas data formats such as TEI, RDF and formats created in relational databases can have different technical properties, they simultaneously share two important characteristics. A first shared characteristic of annotations is that they are typically based on explicit ontologies. According to Davis et al., the creation of a model always implies a set of "ontological commitments", as it demands "a set of decisions about how and what to see in the world".[338] A surrogate is based on decisions concerning the aspects that must be included and the aspects that can be ignored. In the case of plain texts and images, this ontology is implicit. There is no formal document which defines the aspects of the source that are represented. The ontologies that underlie such objective representations can be reconstructed by carefully comparing properties of the model with the properties of the original. Structured annotations, by contrast, serve as explicit manifestations of latent textual phenomena, and are often created to allow for systematic processing. The textual phenomena that are studied can only be retrieved reliably if all the instances of such phenomena are consistently marked as such, using a fixed ontology. The conceptualisations that underlie TEI documents, relational databases or RDF statements are often available explicitly, in the form of XML schemas, ERD diagrams or OWL-based ontologies. Such explicit ontology files dictate the phenomena that may be observed within a particular domain, and additionally stipulate the vocabulary that may be used to describe these phenomena. Explicit ontologies most convincingly exemplify John Sowa's explanation that an ontology may be viewed as "a catalog of the types of things that are assumed to exist in a

[337] Chrostoph Schöch, "Big? Smart? Clean? Messy? Data in the Humanities", in: *Journal of Digital Humanities*, 2:3 (2013), n.pag.

[338] R. Davis, H. Shrobe & P. Szolovits, "What Is a Knowledge Representation?".

domain of interest *D* from the perspective of a person who uses the language *L* for the purpose of talking about *D*".[339]

Structured annotations, secondly, are discrete in nature. Contrary to the words in linear texts, the meaning of individual annotations is not determined by the context in which they appear. A specific data value can usually be isolated completely from values that precede or follow that item. The division which is developed in this chapter between plain texts on the one hand and structured annotations on the other can be connected to the distinction which Lev Manovich discusses, in *The Language of New Media*, between narratives and databases. Manovich argues that novels and cinema have firstly "privileged narrative as the key form of cultural expression of the modern age",[340] while computers and the internet have introduced an alternative, non-linear mode of organisation, namely the database. Manovich presented narratives and databases as "natural enemies", as each of them "claims an exclusive right to make meaning out of the world". Rather than viewing database and narrative as two competing forms of expression, however, it seems more productive to regard narratives and databases as two distinct ways of organising information, each developed for a specific purpose. Narrative is the preferred format for human readers, while the database format is mostly needed to allow computers to process data in a systematic and in an efficient manner. In literary informatics, scholars generally convert linear linguistic compositions with a discursive or narrative structure into a database, which is essentially a collection of discontinuous properties and values, which collectively describe particular aspects of the original linear structure.

4.6. Derived data

In the previous section, a broad distinction was introduced between captured data on the one hand and structured and consistent annotations about these texts on the other. Structured annotations consist of explicit labels which may be connected to text fragments or to texts in their entirety, and which can resolve ambiguities in unprocessed plain texts. Both types of data mostly form a means to an end. Bates clarifies that distinct data values may be organised into larger constellations. Aggregations of such values may result in patterns in which "the sum of the elements constitutes something new, a whole with its own distinct qualities".[341] Such analyses are mostly performed to expose noteworthy characteristics of a corpus as a whole, or to identify individual texts with a conspicuously low or a conspicuously high value for a specific metric. Following Kitchin, the resources that

[339] John F. Sowa, *Knowledge Representation: Logical, Philosophical, and Computational Foundations* (Brooks / Cole 1999), p. 492.

[340] Lev Manovich, *The Language of New Media* (Cambridg, Mass.: The MIT Press 2002), p. 218.

[341] Marcia J. Bates, "Information and Knowledge: An Evolutionary Framework for Information Science", n.pag.

result from further manipulations of captured data or of structured data can be referred to as “derived data”.[342] In this study, a distinction is made between two forms of derived data: (1) summations and (2) processed data. Summation, first, is a basic operation, consisting simply of a count of all the units which can be identified within the captured data or within the structured annotations. Word frequencies form an important example of such derived data. Summations can then be subjected to various forms of statistical learning techniques, such as clustering, counting or filtering. The numbers, lists or patterns that result from statistical analyses are described using the term “processed data”.

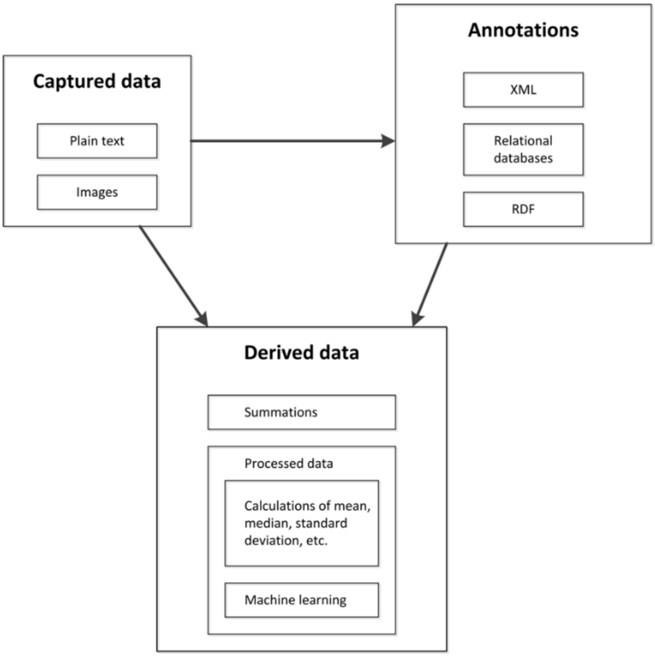


Figure 4.3. Three types research data, which can be analysed or created by three types of algorithms.

In this chapter, three types of research data have been demarcated. Three attendant types of algorithms may be distinguished. A first class of algorithms operates on unprocessed plain texts or images and aims to extract structured annotations which explicitly and consistently describe textual features which are ambiguous or implicit in the original works. This broad activity can be referred to as

[342] Rob Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*, pp. 7–8.

"data creation". A second class of algorithms takes these structured annotations as a starting point, and manipulates these statistically in order to produce derived data. This stage in the research process can be termed "data analysis". Thirdly, it is also possible to perform statistical analyses on captured data directly. In such cases, a single, more extensive algorithm combines the two tasks that can be clearly separated when statistical analyses are based on explicit secondary data. This third type of processing similarly begins with a data creation process. The results of the tokenisation process are stored temporarily, however, in a format that is secluded from the scholar who performs the text mining operation. Once individual textual units have been identified, they can also be analysed quantitatively. The fact that the results of the initial data creation process are not available separately can in some cases be a disadvantage. As procedures for the creation of structured data are often flawed, it is desirable to offer scholars the possibility to manually correct errors in the data that were generated algorithmically. Such interventions can improve the results of text mining processes. If the stages of data creation and data analysis follow one another immediately, there is a risk that analyses are based on erroneous data.

4.7. Conclusion

It has frequently been claimed that scholarship and science are currently being transformed by dint of the seemingly continuous advances in storage capacities and in database technologies. It is probable that the copious availability of machine-readable texts will also affect the methodology and the epistemology of literary studies. Literary informatics entails a form of research in which texts are not necessarily read fully by human scholars, and in which this task can be relegated, partly or wholly, to the computer. In a first phase of the research, digital surrogates need to be acquired of the literary works that are to be studied. At present, literary scholars have access to vast quantities of such digital surrogates, mostly as a result of the mass-digitisation programmes. Such projects normally model the observable aspects of original works via the creation of images or plain texts. In this chapter, such surrogates have been referred to as "captured data". Plain texts retain the inconsistencies and the linear nature of primary sources. Structured annotations, by contrast, result from a rigorous transformation of these resources, in which all aspects that are investigated have been classified unequivocally. These enrichments are necessary because of "the tension between the fierce formalism of code and the inexactitude of human practices and of natural language".[343] Structured annotations aim to transform capricious and ephemeral phenomena into tangible data values which can be processed systematically. they

[343] Caroline Basset, "Canonicalism and the Computational Turn", in: David Berry (ed.), *Understanding Digital Humaniities*, Basingstoke: Palgrave Macmillan 2012, p. 120.

are indispensable in studies which are based centrally on analyses of such phenomena. Statistical analyses of captured data and of annotations result in derived data.

The two broad phases that were distinguished within literary informatics - data creation and data analysis – can be clarified further using the concept of scholarly primitives which was first discussed by John Unsworth, and which has since been elaborated by a number of other scholars. Unsworth uses the term to refer to the "basic functions common to scholarly activity across disciplines, over time, and independent of theoretical orientation".[344] They entail the basic forms of interaction with primary sources which result in an initial set of ideas about these objects. Unsworth argues that "discovery", "annotating", "comparing", "referring" and "sampling", "illustrating" and "representing" form the crucial scholarly primitives. Using Unsworth's concept as inspiration, Palmer et al. have similarly proposed an enumeration of the central activities that are common across academic disciplines. Palmer et al. make a distinction between core scholarly activities such as "searching", "reading", "writing" and "collaborating" on the one hand, and the more specific and discrete scholarly primitives which support these core activities on the other. The latter level includes basic acts such as "chaining", "browsing", "scanning" and "rereading".[345] Blanke and Hedges suggest a list of primitives which is more cognate to Unsworth's original explanation of the term. The authors claim that "discovery", "comparison", "delivery" and "collecting" form the cardinal activities within humanistic research.[346]

It may be claimed that the methods that are employed in literary informatics primarily provide support for annotation, comparison and discovery. Data creation, as discussed in this chapter, basically implies the process of creating annotations. Annotation refers to a process in which objects or fragments within objects are associated with particular descriptive texts. Applications which supply POS tags, lemmas or phonetic transcriptions enrich the bare tokens which are found in the original text with explicit and standardised values which allow for more systematic analyses, and such enrichments can, for this reason, be understood as annotations.[347] Data analysis, second, consists of a description of the differences and the similarities between two or more objects, and can, for this reason, be linked conceptually to the primitive which Unsworth refers to as comparison. This

[344] John Unsworth, "Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?".

[345] Carole L. Palmer, Lauren C. Teffeau & Carrie M. Pirmann, *Scholarly Information Practices in the Online Environment: Themes from the Literature and Implications for Library Service Development.* (2009).

[346] Tobias Blanke & Mark Hedges, "Scholarly Primitives: Building Institutional Infrastructure for Humanities E-Science", in: *Future Generation Computer Systems*, 29:2 (February 2013).

[347] Annotation is based, in turn, on selection, which, according to Unsworth, entails both the identification of objects of interest and the identification of regions or fragments of interest within these objects. An annotation consists of a descriptive term which targets a particular text fragment.

core analytic function consists, more concretely, of the exposition of distributions, correlations or clusters. A distribution graph, for instance, enables scholars to compare the occurrences of particular literary devices in different stages of an author's literary career. Grouping and clustering operations such as k-means clustering or PCA similarly clarify the formal differences between the texts within a corpus.[348] While the results of comparisons may incidentally reveal relevant aspects in themselves, they can also form the bedrock of the scholarly primitive which Unsworth refers to as discovery. This third primitive encompasses the fortuitous identification of a document or of a text fragment whose distinctive or conspicuous characteristics warrant a closer examination. Unsworth stresses that discovery generally has a serendipitous aspect, as the process helps us to locate texts that can "become important to our work in ways that we would not have predicted, and therefore could not have sought".[349]

The discussion of the various types of data may help to clarify the nature of the term "big data". Many discussions of the nature of data in literary informatics reserve the term "big data" exclusively for collections of plain texts. In his discussion of "smart data" and "big data", Christoph Schöch maintains that the latter term refers to machine-readable texts produced by OCR software in digitisation projects, while the former term denotes data which have been created manually. In Schöch's view, smart data are prototypically represented by digital scholarly editions produced on the basis of TEI.[350] Julia Flanders and Matthew Jockers, in a discussion of the conflict between analyses on a micro-level and a macro-level, assume similarly that large-scale analyses of corpora typically take place on the basis of plain machine-readable texts, and that secondary data about the various phenomena that coalesce beyond the lexical codes can only be studied in smaller collections of manually encoded texts.[351] These readings of the term "big data" imply that data expatiating aspects that are implicit in the primary texts can only be small, and that big data are necessarily unstructured. This thesis offers an alternative view, however. In the case study that is conducted as part of this study, it has been shown that data about the linguistic and literary aspects of texts can be supplied both by human encoders and via text mining applications. When the

[348] Unsworth uses the somewhat antiquated term "sampling" to refer to the "result of selection according to a criterion". The process includes "the ability to show distribution and clustering". While Unworth does not explicitly discuss the differences between comparison and sampling, it may be argued that sampling implies a more specific form of comparison. The fragments which are selected through the process of sampling enables scholars to describe differences and similarities between these fragments. See John Unsworth, "Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?", n.pag.

[349] Ibid., n.pag.

[350] Chrostoph Schöch, "Big? Smart? Clean? Messy? Data in the Humanities", n.pag.

[351] Matthew Jockers & Julia Flanders, *A Matter of Scale. Keynote Lecture from the Boston Area Days of Digital Humanities Conference. Northeastern University, Boston, MA. March 18, 2013*, pp. 5–7.

phenomena under investigation can be detected algorithmically, it also becomes possible to create big collections of smart data.

Chapter 5

Machine reading and close reading

5.1. Introduction

Building on from the descriptions of the nature of literary informatics research and of the type of data that are used and produced within such research, this chapter identifies a number of essential qualities of machine reading. The qualities that are established in this chapter allow for a contrastive comparison of algorithmic criticism and traditional criticism based on close reading. This comparison is complicated, however, because of two reasons. A first difficulty is that the concrete possibilities that are offered by machine reading are not fixed. New technological advances in fields such as text mining, natural language processing, computational linguistics and data science often result in turn in innovative possibilities for literary informatics research. Second, the many technological affordances that are available at a given moment in time are never utilised exhaustively. Digital methods are generally adopted only when researchers can imagine relevant applications. Technologies often need to be moulded to specific scholarly requirements, and the efficacy of this process often depends on the inventiveness and the technical proficiency of individual scholars. It is important, for this reason, to make a distinction between the general technological possibilities which are created by text analysis algorithms on the one hand, and the actual ways in which these algorithms have been applied on the other. In assessing the differences between machine reading and close reading, this chapter concentrates principally on general trends in the actual ways in which machine reading has been implemented in past and current research projects. When this text signals specific shortcomings, this is not necessarily a remonstration against machine reading *per se*. In some cases, such deficiencies can be remedied in future studies through a different use of existing technical possibilities.[352]

[352] An additional difficulty is posed by the fact that the properties of close reading are similarly unstable. To allow for a systematic analysis, this chapter uses the definition that was provided in Chapter 2. To reiterate, close reading was defined as a form of engagement which is concerned with the text as an independent unit, which illuminates the meaning of the text though an examination of its form, and which often overlooks the historical and social contexts.

5.2. Distinctive characteristics of machine reading

5.2.1. A focus on linguistic aspects

Chapter 2 explained that the descriptive analysis performed during close reading may focus on prosodic aspects (rhyme, metre and rhythm), devices based on sound (e.g. alliteration, assonance and consonance), devices based on a change in meaning (e.g. metaphor, simile, personification), devises based on ordering or repetition (e.g. anaphora, chiasmus) and on textual phenomena such as diction, mood, tone, volume, texture and intensity. At present, the tools that have been applied and developed within literary informatics research offer limited support for the analysis of these textual phenomena. As was shown in Chapter 3, machine reading is based on algorithms which can recognise and quantify individual words, grammatical and syntactic categories and, to some extent, the semantic contents of texts. To align computer-based research more closely with traditional forms of literary research, it is necessary to develop ways of quantifying those phenomena which are studied in conventional research but which, so far, have often been neglected in computer-assisted research. To quantify a phenomenon, it is important, firstly, to ensure that the computer can recognise instances of this phenomenon. Once such instances have been detected, they can also be counted and analysed statistically.

In some cases, the literary techniques which have been studied in more conventional approaches can be quantified by making use of the basic data that can be generated using existing tools. Data about word frequencies, for instance, can potentially be used to support an investigation of the diction.[353] In characterising the diction of a literary text, it is useful to determine whether or not the author draws from particular registers of speech (e.g. colloquial versus formal, concrete versus abstract, Germanic versus Romanic). Existing tools, however, do not supply any supplementary data that may be used to classify the diction. If such a functionality is needed, scholars will need to manufacture such classifiers themselves, potentially by building on existing tools or lexicons. Data about syntactical categories, which can be generated by POS taggers, can likewise be used to categorise the text's syntax. It can be relevant to classify the syntax either as simple or complicated, or to identify particular syntactic constructions. Such tasks would demand a more advanced processing of the basic annotations that are supplied by POS taggers.

Metre and rhyme can be explored, to some extent, by making use of pronunciation dictionaries. This approach has not been researched extensively, however. David Kaplan has examined the possibility to automate prosodic analy-

[353] As was explained in Chapter 2, the term "diction" refers to the vocabulary or the register of speech that is chosen to express a particular message.

ses, but a number of important issues, such as the strong connection between the rhythm and the meaning of the verse line, and the difficulties caused by diachronic and synchronic variations in pronunciation, have been left unaddressed. By the same token, no standardised tools are available for the detection of devices based on repetition or on unusual word order, such as paronymy or anaphora. It can be conceived, nevertheless, that tools for the detection of literary techniques based on word order or on repetitions of words can be developed when data are available about separate words and about the lemmas of these words.

Next to the literary techniques which can potentially be investigated via algorithms, there are also a number of literary techniques whose detection, in all likelihood, will continue to resist automation. The description of phenomena such as metaphor, personification, mood and toon critically demand an apprehension of the complex semantic environment in which words are used. A computer-based analysis of the meaning of literary texts is complicated for a variety of reasons. Machine reading is premised on the idea that language is predictable, or that the full variety of phenomena can be captured in comprehensive lists. Software applications can be instructed to process signs that are meaningful to human readers, but, like the English-speaking person from John Searle's influential Chinese Room Argument,[354] the machine completely lacks an understanding of what these signs signify. The signification of words can rarely be deduced automatically, since there are no logical connections between words and their denotations. Words have only come to be associated with a specific object or concept through social or cultural conventions. Software applications which aim to assay the semantic contents of texts, such as semantic taggers, often make use of lists which supply possible dictionary meanings. The many distinct social contexts in which words have been used have often resulted in a wide range of potential meanings, however, and the precise signification of polysemous words can, in many cases, be inferred solely by considering the semantic context in which these occur. For domains in which the terminology is relatively stable, researchers in the field of artificial intelligence and natural language processing have defined rules,

[354] Damper offers a concise outline of the argument: “Searle envisages a situation in which he is hidden in a room and is presented questions in Chinese written on an ‘input’ card, posted in to his room by unseen enquirers. Searle knows no Chinese; indeed, he is quite unaware of the enterprise in which he is engaged and is ignorant of the fact that the strange marks on the cards represent questions framed in Chinese. He consults a manual telling him (in English) precisely what equally strange marks to write on an ‘output’ card, which he posts back to the outside world. By virtue of the ‘machine intelligence’ embodied in the manual (which is actually a formalisation of the steps in an AI program), these marks on the output card constitute an answer to any input question. To a Chinese speaker external to the room, by virtue of its question answering ability, the system passes the Turing test for machine intelligence (Turing, 1950), yet the system implemented by Searle-in-the-room is entirely without understanding simply because Searle understands nothing”. See Robert I. Damper, “The Logic of Searle’s Chinese Room Argument”, in: *Minds and Machines, 16:2 (18 October 2006)*, p. 164.

often based on probability statistics, which stipulate that if groups of words occur in particular combinations the text must be about a specific topic.

Such applications cannot easily be developed for the description of the meaning of the signifiers used in literary works. An important obstacle is formed by the fact that words are often used in a figurative sense. They are deliberately given new meanings which are dissimilar from their dictionary definitions. Literary authors, moreover, often experiment with the connotations of texts. Unlike denotations, such connotations are not formally codified. Brooks writes that the paradoxes in literary writing generally "spring from the very nature of the poet's language" which is "a language in which the connotations play as great a part as the denotations".[355] At present, however, no reliable tools are available for finding the connotations of words. Because of the fact that algorithms generally fail to apprehend the connotations of words, it is often difficult to detect instances of irony, paradox and ambiguity, which centrally preoccupied the New Critics.

While human scholars may comment on all aspects of literary texts, the breath of computer-based annotation remains limited to those aspects which can be quantified by algorithmic means. Creating instructions for the recognition of devices based on shifts in meaning, such as metaphor, personification, euphemism, circumlocution, synaesthesia, understatement, authorial intrusion and metonymy is likely to remain cumbersome. An algorithmic resolution of the "unsayable subtlety and stubborn particularity of poetic language",[356] independent of any human intervention, seems beyond the reach of most of today's text mining applications.

5.2.2. Abstracted renditions of collections

The form of close reading that was sanctioned by the New Critics typically concentrates on patterns and relations that are situated at the micro-level of texts, consisting of sentences, paragraphs or stanzas. The New Critical inclination to view works of literature as "well wrought urns, that is, united, cohesive units" also prompted a reluctance to explain textual qualities via references to historical or biographical factors.[357] This narrow and apolitical stance of the New Critics was contested fiercely by theorists associated with post-structuralism and, notably, with New Historicism. Jane Gallop stresses that these movements served as necessary course correctives within literary studies.[358]

Digital methods clearly enable scholars to shift the focus from the micro-level to the macro-level, and to study aspects of collections in their entirety. Margaret Masterman, writing in 1963, surmised that the computer can function as a "tele-

[355] Cleanth Brooks, *The Well Wrought Urn: Studies in the Structure of Poetry*, pp. 5–6.
[356] Willard McCarty, *Humanities Computing*, p. 6.
[357] Verena Theile, "New Formalism(s): A Prologue", p. x.
[358] Jane Gallop, "The Historicization of Literary Studies and the Fate of Close Reading".

scope to the mind", enabling scholars to make new types of observations.[359] She envisioned applications which can lead to radically new insights about the phenomena which are investigated. Bolstered by Masterman's metaphor, Willard McCarty argues that computation is valuable particularly if it can veritably effectuate an epistemological transformation, and if it can convincingly result in "different ideas rather than simply more evidence, obtained faster and more easily in greater abundance".[360] On a more perspicuous level, the image of the telescope is also relevant because of the fact that it accentuates the possibility to expand the scope of literary analyses. Machine reading enables scholars to investigate all the literary output of a specific author, all the texts in a literary genre or all the texts from a particular historical period. This latter possibility clearly differentiates machine reading from close reading. The various forms of derived data about texts collections, which generally result from processes such as such as filtering, sorting or calculation, are difficult to obtain through manual means. The retention of all relevant textual phenomena normally exceeds the mnemonic capabilities of individual scholars. Calculations of the ratio between types and tokens, or of correlations between the frequencies of specific words, would demand a superhuman patience and perseverance. Human critics may admittedly form a global impression of the distribution of specific phenomena via an extensive reading of an author's works, but a computational analysis can frequently modify or subvert such perfunctory impressions by dint of its comprehensiveness and its consistency.

If texts are analysed at the macro-level, this form of research is comparable, in conceptual terms at least, to the methods that were followed by many structuralist literary critics. Inspired by the linguistic theory of Ferdinand de Saussure, structuralist criticism is based on systematic analyses of the language that is employed in literary texts. It typically aims to contribute to an understanding of the inner laws of the style of a literary genre, or of literary language in general. As in literary informatics, the reading focuses on "discourses beyond the limit of the sentence". [361] Studies which concentrate on genres or periods in their entirety cannot equitably be accused of the elitism and the myopia that is often associated with New Criticism, whose aesthetic criteria largely excluded works by female authors, or works produced in developing countries. The rules that are implemented in algorithms can usually be applied to any machine-readable text, without discrimination. The texts to be mined obviously need to be available in a machine-readable form. While the methods in themselves are not partial to particular types

[359] Margaret Masterman, "The Intellect's New Eye", in: *Freeing the Mind: Articles and Letters from The Times Literary Supplement during March-June 1962*, London: Times Publishing Company Ltd. 1962.

[360] Willard McCarty, "A Telescope for the Mind?", in: Matthew Gold (ed.), *Debates in the Digital Humanities*, Minneapolis: University of Minnesota Press 2012, p. 114.

[361] Michael Groden, *The Johns Hopkins Guide to Literary Theory and Criticism* (Baltimore: Johns Hopkins University Press 1994), pp. 890–891.

of texts, or to particular types of authors, it may be assumed that the nature of research corpora can also be determined, to some extent, by the policies of digitisation programmes, which may occasionally reflect a bias. While certain digitisation initiatives indeed limit themselves to texts which are considered part of a certain canon, the majority of programmes include institutional holdings or on historical periods in their entirety, however, without discriminating on the basis of the contents of texts.

Machine reading can be applied productively to study questions of literary history. Visual representations of derived data can astutely enable scholars to investigate historical developments in phenomena such as literary genres or literary productivity. One aspect of machine reading which may potentially undermine its effectiveness for historical research, nevertheless, is the fact that it generally treats all texts equally, even when they originate from different historical era. In this respect, literary informatics reiterates the ahistoricism of New Criticism. Computational analyses are often based exclusively on the words of the texts, disregarding the political and social contexts of literary works. Although the New Historicist movement has drawn attention to the notion that the form of a particular text can only be understood properly by considering the historical context, digital tools generally apply the exact same algorithms to all the texts in a corpus.[362] The consistency with which machine reading algorithms analyse texts implies a return to the precept that literary texts ought to be viewed as authorless and timeless documents.

As has been shown, however, the possibility to expand the scale can be very beneficial to research in the field of literary history. An important requirement, clearly, is that the metadata associated with the various texts must include an indication of the date of creation. When it is estimated that the criteria for detecting specific formal features are accurate and sufficiently inclusive, machine reading may disclose historical developments in the occurrences of these features. Literary informatics is simultaneously a formalist approach and a method which can be used to expose historical patterns. It allows for a detailed examination of the language that is used with a text, and it also enables scholars to place such characteristics within a broader historical context. Because of this two-fold attentiveness, computer-assisted research appears to have some allegiance to the emerging field of New Formalism. The latter field sanctions a form of close reading which is informed by a historical awareness. It recognises “the form literature has taken and the aesthetics it has appropriated”, and uses a knowledge of literary history to explain what makes a particular text distinctive or prototypical.[363]

[362] The fact that word meanings can change over the course of history produces clear challenges for studies which focus on diachronic developments. Text analysis tools can technically be instructed to treat these words differently, based on the dates of creation of the texts in which they are used. Such applications are rare, however.

[363] Verena Theile, “New Formalism(s): A Prologue”, p. 5.

5.2.3. Abstracted renditions of individual texts

While literary history mostly investigates literature as a collective system in an attempt to extract general laws, the discipline of literary criticism typically aims to expose the unique properties of extraordinary works of literature. For critics, algorithm-based reading can be useful only if it can actually reveal meaningful aspects of the texts they are interested in. Machine reading can be relevant to literary criticism in two distinct ways. Analyses at the macro-level, first, can help scholars to discover individual texts with noteworthy properties. They may reveal, for instance, that a specific cluster of texts has exceptional values for a given metric, or they may indicate, conversely, that texts which were traditionally considered exceptional appear to be completely ordinary when viewed statistically. Such findings can stimulate reflections about texts, and such reflections can in turn spur new interpretations. Computational analyses may result in patterns which, in many cases, can only be explained by revisiting the individual texts. As was also shown in Chapter 4, studies which use digital methods for the purpose of literary criticism often use the results of the structural analyses of the data as a starting point for subsequent qualitative analyses performed by the human scholar. The scholars who were associated with the MONK project, for instance, view text mining primarily as a method which can initiate critical provocations.[364] In such studies which merge statistics and hermeneutics, the close reading can verify or falsify the results that are produced by machine reading, and vice versa.

Next to focusing on large collections, the computer can also concentrate on the infinitesimal details of individual texts. All computational analyses initially derive from prior descriptions of the minutiae of text fragments. By combining different algorithms for the identification of words, syntactical categories or literary devices, scholars can often collect more details than would ever be possible via conventional close reading. Such atomic observations at the micro-level can subsequently be aggregated at many different levels of analyses, to reveal patterns that lie beyond these individual textual units. Digital text analysis tools can serve both as microscopes and as macroscopes, as they can focus on any level of analysis in between the massive and the minuscule.

When data are shown at the level of individual works or at the level of smaller fragments within these works, such perspectives can enable the form of research which Alan Liu refers to as "close reading 2.0". Liu observes that the digital humanities have concentrated predominantly on the exploration of big data collections, and argues that the field has undervalued computer-assisted analyses of "individual objects of humanistic interest in the era of distant and macro-analytics".[365] A

[364] Matthew Kirschenbaum, "The Remaking of Reading: Data Mining and the Digital Humanities".

[365] A. Liu, "The State of the Digital Humanities: A Report and a Critique", in: *Arts and Humanities in Higher Education*, 11:1-2 (1 December 2011).

number of scholars, nonetheless, have used digital methods to create an improved understanding of singular works. Tanya Clement's algorithmic analysis of Gertrude Stein's novel *The Making of America*, for instance, is often cited as a highly innovative and a strongly compelling illustration of the potential of computational methods. In his monograph *Reading Machines*, Stephen Ramsay discusses a systematic investigation of the stylistic differences between the six speakers in the novel *The Waves* by Virginia Woolf, and Eric Bulson has used digital methods to examine the "numerical unconscious" within James Joyce's *Ulysses*.[366] Such experiments with text mining and machine learning are primarily driven by the conviction that computation can generate innovative and surprising perspectives on texts which have already been subjected to minute examination via conventional close reading at an earlier stage.

A crucial quality of machine reading is that it enables scholars to produce systematic abstractions of texts. Literary texts often contain complicated combinations and repetitions of words, literary devices and connotations. Via algorithmic analysis, scholars can partly reduce this complexity and focus closely on a limited number of textual aspects. Stephen Ramsay places algorithmic criticism in a much broader context and argues that all criticism, based either on digital or on analogue resources, is essentially algorithmic in nature. Critics invariably study texts from a particular perspective, and the construction of such critical angles "relies on a heuristic of radical transformation". Criticism entails the creation of "a new text in which the data has been paraphrased, elaborated, selected, truncated and transduced".[367] When scholars view texts through a critical lens, they accentuate and magnify specific aspects and obscure certain other aspects. The transformations that can be created via computation typically differ in the sense that they are generally based on logical or mathematical operations such as classification, filtering or clustering.

5.2.4. Non-responsive and context-independent analysis

In the case of human reading, the interaction with the text is mostly of a responsive and flexible nature. Readers recognise that the meaning of a particular word can be affected by factors such as religion, gender and social status, and they tend to apply their knowledge of the social and the historical origin of texts during their assessments of particular text fragments. Following Gadamer, it can be posited that interpreters generally enter into a dialectic relation with texts, in which prior conceptions of the nature of the work can be modified during the reading process.[368] During descriptive analyses of texts, human scholars commonly make use of certain rules for the recognition of literary devices, but they also permit deviations

[366] Eric Bulson, "Ulysses by Numbers", in: *Representations*, 127 (2014).

[367] Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism*, p. 16.

[368] David Hoy, *The Critical Circle: Literature, History, and Philosophical Hermeneutics*, p. 70.

from such rules in unclear of ambiguous cases. The computer, by contrast, simply applies the rules that are stipulated in an algorithm in a staunch and unwavering manner. The machine reading process is unrelentingly consistent, and produces the exact same type of metrics for all the texts in a particular corpus. It is centrally based on counts, and the rules that stipulate the criteria for being counted allow little room for exceptions.

Additionally, a close reading of a text often consists of a minute examination of the various phenomena that occur on a micro-level. Scholars traditionally consider the way in which the effects of specific literary techniques may strengthen or, perhaps, undermine, the effects of other literary techniques within the same passage. Such investigations of the many complicated connections that can exist between words and literary devices can help to illuminate the way in which the text produces meaning. Whereas conventional close reading is mostly attentive to the total effect that is produced by the various literary techniques in combination, many of the computational analyses which have been surveyed in Chapter 3 concentrate on singular textual aspects. Studies frequently limit themselves to analyses of most frequent words, or to analyses of syntactic categories, without probing for the potential correlations that may exist between distinct literary techniques.[369] Machine reading, importantly, is based on a form of processing which is context-independent. Text mining algorithms are typically based on simple counts of the occurrences of textual aspects, and once a textual phenomenon has been converted into a number or into a label, it is difficult to use characteristics of the original context during analyses of these numbers. The data values are disoriented from their original setting, and they become entities which can be manipulated on their own terms. Aspects of style are frequently investigated solely through a bag-of-words model, but the unequivocal neglect of the original word order categorically precludes the investigation of what appear to be essential features of a writing style. The style of a particular author can be characterised by a particular timbre, a punctuation regime, unusual word combinations, the use of alliterative effects, and the overall flow and rhythm of sentences, but such stylistic characteristics are usually disregarded. Many of the studies which have examined the differences between literary characters have likewise focused exclusively on the differences between the words that are spoken. Other aspects, such as the development of personalities throughout a text, or distinctions between flat and round characters, cannot readily be quantified and are consequently left out of consi-

[369] There are a number of important exceptions, however. David Kaplan, for instance, quantified a large number of aspects of literary works and examined these simultaneously through various types of multivariate analyses. See David Maxwell Kaplan, *Computational Analysis and Visualized Comparison of Style in American Poetry*. Kaplan's approach was adopted and modified by Justine Kao and Dan Jurafsky. See Justine Kao & Dan Jurafsky, "A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry".

deration. Algorithmic processing often takes phenomena out of their original context and proceeds to scrutinise these in a rather narrow fashion.

5.2.5. Suspension of interpretation

In Chapter 2, it has been explained that the descriptive analyses that are performed as part of close reading processes generally provide the groundwork for interpretation. Within conventional literary criticism, textual analyses traditionally seek to elucidate the meaning or the themes of texts. Interpretation often consists of a consideration of the connection between this meaning and its contents. The thematic concerns of a text cannot unproblematically be established via algorithms, however, as there is rarely a close and predictable relationship between the words that are used and the themes which are being developed by these words. Some of the studies that have been discussed in Chapter 3 have attempted to identify themes computationally, however, using techniques such as topic modelling or semantic tagging. Arguably, the "theme" has been defined somewhat narrowly in such studies as the literal "aboutness". Methods based on vocabulary may disclose the setting of a text, the images which are evoked, or the objects or events which are depicted. In figurative language, however, the referents of words generally stand in a symbolic relation to the text's abstract and more recondite thematic concepts. The identification of such abstract themes, moreover, is often debatable. A degree of intersubjective agreement may be reached concerning the concrete objects and events which are evoked by words, but a description of the more abstract concepts connected to works of literature inevitably demands a subjective interpretation.

In Chapter 2, it was argued additionally that the validity of an interpretation cannot be considered independently from the interpreter. In view of Gadamer's hermeneutical philosophy, an interpretation can be considered valid if it results in an improved self-understanding on the part of the reader. The attempts to grasp the meaning of the text and to clarify the concrete ways in which the author produces and reinforces this meaning via form and via language eventually serve to generate new insights and new questions about man's experience of the world. The interpretation of poetry demands a recognition of the fact that texts can make meaningful statements about human experience. Brooks and Penn explain that poetry is "a response to, and an evaluation of, our experience of the objective, bustling world and of our ideas about it" and that it is concerned with "the world responded to sensorially, emotionally, and intellectually".[370] Since the precise ways in which a text produces meaning is likely to remain incommensurable, text analysis tools cannot support unsupervised explorations of the relation between form and meaning.

[370] Cleanth Brooks, *Understanding Poetry*, p. 9.

More broadly, quantitative analyses of data on linguistic features do not directly generate clear-cut answers about the meaning of the text. They initially result in statistical products which still need to converted into statements about the domain which is investigated. Computers may be used to expose patterns in the occurrences of specific features, but they cannot independently account for the nature of these patterns. The computer can augment "the critic's power of perception and recall in concert with conventional principles",[371] but the human critic continues to bear the responsibility to supply a logically coherent interpretation, based on the patterns which are suggested by quantitative analyses. Craig stresses that any "departure from the purely enumerative ... is an act of judgement and is open to question".[372] To clarify or explain the trends or the patterns that emerge from such forms of processing, scholars will often need to make certain conjectures, making use of additional, extra-textual information. To explain the significance or the relevance of such findings, a leap is needed from a purely quantitative approach to a more interpretative engagement.

It is misleading, nevertheless, to claim that computational analyses are fully devoid of interpretation. As will be discussed more elaborately in the following chapters, the data that are analysed by text mining algorithms typically result from subjective and debatable decisions about the way in which complicated humanistic phenomena ought to be quantified. The results that are produced by algorithms often vary strongly along with the initial parameters that are provided and with the more specific settings that are chosen. The nature of analytic processes can consequently reflect idiosyncratic preferences and subjective interpretations of the goal and the scope of the analysis. In claiming that algorithmic processing suspends interpretation, I mainly aim to stress the notion that quantitative analyses do not provide any explicit explanations of their results. They basically offer descriptive information about the objects which are investigated, and they do not in themselves indicate the scholarly relevance of particular findings. Interpretation is likely to remain an inherently human capacity.

In Chapter 2, it was also explained that literary critics often aim to evaluate literary texts in a qualitative sense. Among the many decisions which algorithmic criticism aims to automate, the determination whether or not a text has literary quality will probably continue to be cumbersome. Works of literature have been judged in many different ways, and the methods which have been used to assess texts in an evaluative manner, moreover, are often difficult to formalise. In their monograph *Theory of Literature*, Wellek and Warren explain that the different schools of literary criticism have spawned a variety of evaluative norms and standards. It has been claimed, for instance, that great literature ought to be

[371] John B. Smith, "Computer Criticism", p. 14.
[372] Hugh Craig, "Stylistic Analysis and Authorship Studies", p. 277.

"expressive of the 'great' values of life".[373] Alternatively, Russian Formalists such as Victor Schlovsky and Roman Jakobson have evaluated the quality of literature by considering the level of estrangement from mundane language.[374] Wellek and Warren note that this latter criterion is ultimately relativist, as later readers can clearly grow accustomed to modes of expression which were previously innovative. Additionally, the value of a work of literature cannot be judged solely by considering the reactions which are evoked within the reader, as this approach "does not correlate the nature of the response with the nature of the object". Judging a work exclusively on the basis of the poem itself, conversely, presupposes "absolute standards thought of as existing without reference to human need or cognition".[375] All acts of critical evaluation demand a prior definition of the nature and the objectives of literature in general. Since such views on the artistic goals that literary texts should strive to attain are almost inevitably tied to particular approaches in critical theory, it seems impossible to determine literary quality in an absolute sense. In most cases, the outcomes of quantitative processing still need to be assessed on a qualitative level by a human scholar, who can ascertain if a text can indeed function as "a cause, or a potential cause, of the reader's 'poetic experience'".[376] For a machine that requires predictability and formalisation, it is surpassingly difficult to come to grips with the notion that literary quality can emerge from unanticipated deviations from bendable aesthetic norms.

5.3. Conclusion

Although its nature may evolve as technology progresses, algorithmic criticism is presently a formalist critical approach, concentrating principally on the vocabulary and the grammar of literary works. Because of the current inability to reliably describe the complex semantic contexts of tokens, algorithms can currently only identify a limited number of literary or linguistic phenomena. Algorithms, importantly, do not add any interpretation, but they can be used to produce statistical artefacts which may provoke interpretation. Whereas the technology expedites the analysis of non-exclusive and non-elitist corpora, the sizes of the corpora often remain modest because of the fact that research frequently blends statistical processing with a hermeneutic and a qualitative form of engagement.

Literary scholars aiming to adopt computational methods currently encounter difficulties and limitations which may compromise the value of these methods. For this reason, it is important to investigate if it is possible to address or to remedy some of the central challenges that have arisen within literary informatics. Four

[373] Rene Wellek & Austen Warren, *Theory of Literature* (Harmondsworth: Penguin Books 1963), p. 245.
[374] Ibid., p. 242.
[375] Ibid., p. 248.
[376] Ibid., p. 249.

important areas of further research may be identified. First, it can be useful to examine the feasibility and the desirability of developing new algorithms for the detection of the literary devices which are commonly studied in conventional literary research. At present, text analysis tools mostly concentrate on vocabulary or on syntactical categories, and often fail to create data on the occurrences of literary devices based on sonic patterns or devices based on changes in meaning. Second, if it is found that it is indeed possible for algorithms to detect some of the literary devices that have hitherto been neglected in computer-based scholarship, it is equally important to reflect on the interpretative possibilities that can emanate from statistical analyses of such data. Data analysis, more generally, aims to extract relevant information from text corpora through the application of a variety of statistical procedures. Data analysis is frequently an onerous process, in which the concrete needs emerging from research questions must be connected to specific analytic methods. Which statistical algorithms can genuinely produce new and relevant insights about the literary works that are studied?

A third aspect which deserves closer scrutiny is the fact that it is based on data whose formats invariably imply ontological commitments. Such predefined ontologies can limit the ways in which the data can be analysed. Does the imperative of having to work with data formats obviate particular types of questions, or are scholars still free to study the questions they are genuinely interested in? When scholars manage to develop new algorithms for the detection of specific literary devices, can such observations still be captured using existing data formats? Fourth, additional research needs to be conducted into the scholarly possibilities of data visualisations. Graphical rendition of large volumes of quantitative data can be valuable, as they often allow for a swift identification both of values which are commonplace and of values which are exceptional. Since literary scholars have rarely been trained in the creation of data visualisations, however, it can occasionally be difficult to read and to interpret such non-textual resources correctly. An additional difficulty is that data visualisations often result in abstract patterns. Since literary criticism aims to demonstrate the particular value of individual works of literature, the question may be asked if such abstractions can veritably be of value to literary scholars.

For the purpose of this thesis, these four sets of questions have partly been investigated on a practical level during a case study. This case study was conducted to supplement the results of the theoretical examination of the strengths and the shortcomings of literary informatics with insights emerging from hands-on experimentation with computational methods. As has been argued by various authors, hands-on experiences are often indispensable in studies that seek to understand the ramifications of the digital medium. Practical work often produces concrete challenges which could not have been anticipated by a purely theoretical framing of the subject. Writing about digital humanities research, McGann explains that digital applications may usefully be viewed as tools for “imagining what we don't know”.[377] Experiments with computation often encourage scholars to develop knowledge about problems which had initially been outside of their awareness. The practical experiments that were conducted for this thesis usefully helped to produce a more solid understanding of the nature of algorithmic criticism, and, as will be shown in the following four chapters, they effectively helped to trace some of the crucial difficulties connected to the creation, the representation, the analysis and the visualisation of data about literary texts.

[377] Jerome McGann, *Radiant Textuality: Literature after the World Wide Web* (New York: Palgrave Macmillan 2004), pp. 105–120.

Chapter 6

Data creation

6.1. Introduction

The previous chapter concentrated on the question whether or not computational methods can beneficially aid scholars in unravelling the manifold linguistic and rhetorical aspects that can give literary works their unique quality. The aspiration to automate some of the core activities within the context of literary criticism generally stems from the expectation that digitisation can yield particular advantages with respect to speed, precision, or quality. This impulse to digitise is based additionally on the conviction that particular core activities are indeed amenable to digitisation. In his widely cited essay "Computing Machinery and Intelligence", Alan Turing has stressed that the value of the computer, as a universal machine, lies in the fact that it can "mimic any discrete-state machine",[378] provided that the functions of this machine can be described via a finite number of unequivocal activities. When scholars take "the computational turn",[379] they essentially need to consider if sections of their existing scholarly practices can be formalised through algorithms. Research projects in the field of literary informatics aim to address the questions which are asked traditionally within literary studies in innovative ways by making use of computational methods. To ensure that computation can veritably be supportive of the discipline in which the tools are adopted, it is necessary to take the existing practices and traditional interests as the point of departure, and to manipulate the toolset in such a way that it can be used, as much as is possible, in the service of these traditional objectives. Matching the possibilities that are offered by informatics to existing scholarly objectives often demands many efforts and much creativity, however. The digital medium principally provides support for calculations, and it can process data only if these explicit and unequivocal. Because of these demands, it is not always possible to represent conventional scholarly approaches as quantitative operations.

The digital heuristics that emerge eventually are often an amalgamation, resulting from a negotiation between what is desirable from a scholarly perspective on the one hand and the limitations and the affordances of computation on the other. On the one hand, Turing's conceptualisation of computing implies that digital technology is malleable. The universal machine was not built for a unitary

[378] Alan Turing, "Computing Machinery and Intelligence", in: Noah Wardrip-Fruin & Nick Montfort (eds.), *The New Media Reader*, Cambridge Mass.: The MIT Press 2003, p. 54.

[379] Caroline Basset, "Canonicalism and the Computational Turn".

objective, but can be transmuted by its users to make it serve particular purposes. There may also be a backlash, however. While it may be true that we can influence the nature of the technologies we use, these technologies may simultaneously, perhaps unconsciously, shape us.

For the purpose of this thesis, the concrete obstacles in aligning traditional practices and scholarship based on data processing have been investigated in a practical manner, by conducting a case study. The case study consisted, more specifically, of a computer-based analysis of the poetry of the Northern Irish poet Louis MacNeice. Robin Skelton notes that Louis MacNeice is “one of the master craftsmen” in English poetry and also stresses his poems are “essential reading for anyone who cares to study or to practice the intricacies of English verse”.[380] MacNeice’s poetry simultaneously exemplifies the sundry obstacles that may emerge from the manifold interpretability of language. Reflecting on his own verse craft, MacNeice wrote that language is generally “a traffic in symbols” and added that “these symbols are plastic - an endless annoyance to the scientist but God's own gift to the poet”.[381] Interestingly, as will be shown, MacNeice’s main thematic concerns mirror some of the central challenges in the field of literary informatics.

As was indicated in the previous chapter, computational analyses of texts are often based on prior quantifications of low-level linguistic features, such as the most frequent words or occurrences of specific grammatical constructions. Studies which make use of conventional text tools, and which are consequently based on such formal characteristics, are, to some extent, intellectually remote from traditional forms of literary scholarship. In the case of poetry research, the focus is generally on the description and the interpretation of aspects such as meter, figures of speech, imagery or themes. As part of the case study, a basic software application was developed for the automated detection of a large number of literary devices. Unless indicated otherwise, all software has been programmed by myself.[382] Where possible and where relevant, the studies and the functionalities which have been discussed in Chapter 2 have been used as guidelines.

The case study was divided into three phases. First, an analysis was made of a large part of the existing criticism on MacNeice. An inventory was compiled of some of the questions that were asked about the verse, and of the methods that were used to answer these questions. Most pertinently with respect to the case study’s central aims, data was also collected about the various literary devices that have been identified by critics. A second phase focused on the practical obstacles involved in the creation of data about literary devices. The observations which were

[380] Robin Skelton, “Celt and Classicist: The Versecraft of Louis MacNeice”, in: Terence Brown & Alec Reid (eds.), *Time Was Away*, Dublin: Dolmen Press 1974, p. 53.

[381] Louis MacNeice, "Experiences with Images", in Alan Heuser (ed.), *Selected Literary Criticism of Louis MacNeice,* Oxford: Clarendon Press 1987, p. 155.

[382] All the code that was developed for the purpose of this dissertation can be found at <https://github.com/peterverhaar/Phd>

made in the existing criticism formed inspiration for the design of the text analysis software. The case study aimed to investigate whether or not the statements made by critics could also be produced on the basis of text mining tools. In a third phase, the data that were produced were also analysed statistically. One of the objectives in the third phase was to explore if data processing may help to produce new readings and if these can contribute to an improved understanding of the texts. The aim was to design analytic procedures which can help to address the issues which were also raised in the existing criticism. The case study is largely based on the hypothesis that specific aspects can be investigated more thoroughly and more consistently if they are analysed through computation. The case study also investigated whether or not computers can enable scholars to answer existing questions differently, or, more ambitiously, if digital methods can also allow critics to ask entirely new types of questions.

The following section firstly gives a brief summary of the existing criticism. For the largest part, however, section 6.1. discusses the difficulties connected to the creation of data. This chapter discusses the various functionalities that have been programmed in close conjunction with the reasoning that has been following during the implementation of these functionalities. Chapter 7 explains how the data that have been created are stored, focusing, more particularly, on the benefits and the disadvantages of a number of data formats. Chapter 8, finally, focuses on the broader patterns that can be generated via analyses of the resultant data.

6.2. Case Study

6.2.1. Introduction

To clarify the general background of the central case study, this section offers a brief discussion of the life and work of Louis MacNeice, together with a summary of the main issues that are addressed in the various critical studies of MacNeice's work. MacNeice was born in 1907 in Belfast, and grew up in Carrickfergus, where his father was a Church of Ireland rector. MacNeice's mother died when he was five years old. Since the age of ten, he was educated at public schools in England and at Merton College in Oxford. In 1929, MacNeice started working as a lecturer in Classical literature at Birmingham University, and, during this period, he also published his first volumes of poetry. MacNeice became a highly prolific author, writing ten volumes of poetry, a large number of radio plays and a sizable body of literary criticism. During his lifetime, MacNeice was mainly known as a member of the group of thirties poets, which also included W.H. Auden, Cecil Day Lewis and Steven Spender.

After MacNeice's early death in 1963, a large number of critical assessments have been made of his full oeuvre and if its impact.[383] Most critics agree that MacNeice's literary output can broadly be divided into three phases. The poet initially rose to fame in the early 1930s following the publication of his first volumes *Poems* and *The Earth Compels*. The poems in these volumes have frequently been lauded for their strong sensual and visual qualities, the skilful use of meter and rhyming schemes, and the overall colloquial and free-flowing register that is adopted. The long poem *Autumn Journal*, largely written in 1938, is frequently considered MacNeice's masterpiece. The early 1950s, however, during which MacNeice wrote *Ten Burnt Offerings* and *Autumn Sequel*, are generally viewed as a period of creative impasse. About *Ten Burnt Offerings*, Allan Gillis writes that "the words fail to animate, and the verse remains stubbornly flat-lined".[384] In the second half of the 1950s, MacNeice gradually began to develop a new style, in which he experimented with parable and with different types of rhymes. A number of scholars have noted that the last three volumes *Solstices*, *Visitations* and *The Burning Perch* contain some of MacNeice's best poetry. Goodby stresses that, in this third and final phase, the poet managed to achieve a "lyric compression, revealing an adjustment to the darker climates of Cold War and middle age".[385]

Terence Brown argues that MacNeice's verse is informed fundamentally by a profound scepticism, and by a refusal to commit fully to any particular philosophy or religious creed. Brown connects this scepticism to biographical factors, such as the early loss of his mother, and to his state of living as an exiled Irishman in England.[386] The first volumes in particular represent the world of sensual experiences as intrinsically complex and transient, and stress that it is impossible to explain or to understand such plurality through a finite set of values or rules. Derek Mahon stresses similarly that MacNeice's verse reflects a clear distrust of generalisations. Throughout his literary career, however, the poet was beset simultaneously by a fear that a lack of commitment and the consequent superficiality cannot be sufficient, and he continued to hanker after more permanent and more profound moral values. The poetry, in short, displays a clear tension between profundity and superficiality.

[383] Important book-length studies were written by Terence Brown, John Stallworthy and Edna Longley. Brown and Longley have also edited collections of essays, such as *Studies into Louis MacNeice* and *Time was Away*.

[384] Alan Gillis, ""Any Dark Saying": Louis MacNeice in the Nineteen Fifties", in: *Irish University Review*, 42:1 (20 May 2012), p. 108.

[385] John Goodby, "Louis MacNeice", in: David Scott Kastan (ed.), *The Oxford Encyclopedia of British Literature*, Oxford: Oxford University Press, p. 365.

[386] Terence Brown, *Louis MacNeice: Sceptical Vision* (Dublin; New York: Gill and Macmillan 1975), p. 20, and passim.

Various scholars have also noted that MacNeice's basic scepticism and agnosticism had affected his political orientations. While contemporaries such as Auden, Spender and Day-Lewis clearly reacted to the social and the political climate of the 1930s by writing verse that was overtly socialist, Peter McDonald notes, by contrast, that MacNeice's 1935 volume *Poems* seems "aggressively neutral".[387] MacNeice most frequently addresses politics and social commitment in an indirect manner. His main interest was in the issue of "how far the self is able to marginalize the other into mere 'context' and how far it is the context, the other, which gives meaning to the self". MacNeice's leanings, nevertheless, were predominantly left-wing.[388] After his travels to Iceland and to Spain at the onset of the civil war, MacNeice clearly became more politically engaged.[389] Gillis observes that *Autumn Journal* "fused the personal and the communal in a poetics of social awareness and commitment".[390] The disillusionment that followed the Second World War, however, appears to have resulted in a disinterest in political involvement and in misgivings about the value of social commitment. In MacNeice's later work, "the sense of both self and society has become morbidly phantasmagorical".[391]

MacNeice's troubled relation with Ireland has also attracted much critical interest. The poet had left his native country at an early age, and, throughout his lifetime, Ireland remained a place both to admire and to reject. While many poems idealise the West of Ireland, texts such as "Valediction", "Neutrality" and "Autumn Journal" also contain vehement reactions against Irish politics and Irish culture. Importantly, however, MacNeice has had a profound impact on the poetry written in Ulster in the 1970s and the 1980s. Heather Clarke records that, while MacNeice was viewed largely as a minor poet working in the shadow of W.H. Auden during his lifetime, the poet's reputation was restored posthumously by Northern Irish poets such as Seamus Heaney, Derek Mahon and Michael Longley, who viewed MacNeice as a major figure in the poetic tradition of Ulster and who recognised him as a "model of cultural transience and displacement".[392] Clarke also observed, however, that this repatriation was also a reinvention, as these poets needed to ignore MacNeice's political commitment during the Second World War. Derek Mahon writes that, while the poetry is often not specifically about Ireland, MacNeice nevertheless has "some sort of Irish sensibility".[393] John Goodby argues

[387] Peter McDonald, "The Falling Castle", in: Jacqueline Genet & Wynne Hellegouarc'h (eds.), *Studies on Louis MacNeice*, Caen: Presses universitaires de Caen 1988, s. 1.

[388] Goodby notes that MacNeice largely adopted "a pragmatic, skeptical socialism". See John Goodby, "Louis MacNeice".

[389] It is clear from his discussion of the role of poetry in society in the essay *Modern Poetry*, and it is also apparent in a large number of poems in *The Earth Compels*, and, most notably, in *Autumn Journal*.

[390] Alan Gillis, ""Any Dark Saying": Louis MacNeice in the Nineteen Fifties", p. 107.

[391] Ibid., p. 106.

[392] H. Clark, "Revising MacNeice", in: *The Cambridge Quarterly*, 31:1 (2002).

[393] Derek Mahon, "MacNeice in England and Ireland", in: Terence Brown & Alec Reid (eds.), *Time Was Away: The World of Louis MacNeice*, Dublin: Dolmen Press 1974, p. 113.

similarly that MacNeice addressed quintessentially Irish themes, and that his "interest in relativity and flux, his constant attempts to deconstruct binaries, his concern with the self and with tradition as potential self-betrayal make him an exemplary Irish writer".[394]

This study is interested in the algorithmic investigation of literary devices, and, for this reason, it was particularly important to collect critical descriptions of the way in which MacNeice expressed these themes through language. Many critics have commented on the relaxed nature of the meter, the elegance of the rhyming schemes, and the frequent use of devices such as alliteration and assonance. About the use of rhyme, MacNeice argues, in his long essay *Modern Poetry*, that perfect rhymes can add musicality to the verse, while it also suggests an insincerity. As a compromise, poets can deploy variations, such as "internal rhyme, off-rhymes, bad rhymes and 'para-rhymes'" or they can rhyme "a stressed against an unstressed syllable".[395] In his own work, MacNeice experimented with the many similarities in sounds and with various types of rhyme. Skelton stresses that the poet frequently deployed "elaborate patterns of near-rhyme, assonance and consonance",[396] and Longley has similarly commented on the "importance of emphatic rhythmical punctuation — like assonance, internal rhyme and refrain".[397] In a close reading of the poem "Donegal Tryptych", Terence Brown notes that "the use of assonance and alliteration, the sheer vigour of the diction, together with the patterning, all draw attention to the language as an almost physical object".[398]

Various critics have focused on the use of repetitions. Neil Corcoran notes that the poem *Leaving Barra* contains various forms of repetition, such as a reiteration of the exact same word in two consecutive lines, the use of an identical phrase at the beginning and at the end of the poem ("the dazzle on the sea") and a repetition of parts of words in other words on the same line ("the rain and the rainbow"). Next to being aurally pleasing, such repetitions are also "thematically functional",[399] as such repetitive elements are "sensitively mimetic of the mind in progress – self-scrutinising, self-corrective, advancing hesitantly but keeping moving".[400] Allen Gillis observes similarly that MacNeice's poetry frequently contains "repetitive riffs" and that the poet often uses "chiasmus and chiastic-like effects".[401] Interestingly, such chiastic repetitions create a "paradox of movement and stasis", and

[394] John Goodby, "Louis MacNeice".

[395] Louis MacNeice, *Modern Poetry: A Personal Essay*, p. 131.

[396] Robin Skelton, "Celt and Classicist: The Versecraft of Louis MacNeice", p. 43.

[397] Edna Longley, "Louis MacNeice: Aspects of His Aesthetic Theory and Practice", in: Jacqueline Genet & Wynne Hellegouarc'h (eds.), *Studies on Louis MacNeice*, Caen: Presses universitaires de Caen 1988, s. 1.

[398] Terence Brown, *Louis MacNeice: Sceptical Vision*, p. 172.

[399] Neil Corcoran, "The Same Again? Repetition and Refrain in Louis MacNeice", in: *The Cambridge Quarterly*, 38:3 (4 August 2009), p. 216.

[400] Ibid., p. 214.

[401] Alan Gillis, ""Any Dark Saying": Louis MacNeice in the Nineteen Fifties", p. 111.

they convey the way in which things “are both singular and multiple at the same time”.[402]

Robin Skelton, in his essay “Celt and Classicist: the Versecraft of Louis MacNeice”,[403] argues that MacNeice’s Irish background clearly transpired via the usage of poetic devices which are characteristic of Celtic verse. MacNeice often created lines with an uneven number of syllables and alliteration, near-rhyme, assonance and consonance. In a close reading of the poem “Aubade”, Skelton demonstrates that a Celtic influence is noticeable in the use of deibhide rhyme, in which unrhymed syllables rhyme with stressed syllables.[404] There is also a preference for slant rhyme over perfect rhyme, as the poem stresses the acoustic similarities between the words “apple” and “happy” and between “dawn” and “war”. Skelton also discusses various occurrences of internal or Aicill rhyme. A specific case of Aicill rhyme occurs when the consonants of the final word of one line are repeated in the consonants of the first words on a line that follows. This device is used in “Order to View”, in which the consonants in the word “crypt” on line 7 are repeated in the word “empty” on line 9.[405]

The body of literary criticism is extensive, and it is impracticable to include a complete discussion of the debate on the merits and the shortcomings of MacNeice’s writings in this section. This section principally aimed to highlight a number of topics, and to identify a variety of questions that can be addressed further in the case study. Critics of MacNeice’s poetry have focused, in short, on the stylistic and thematic differences between the separate volumes, the impact of MacNeice’s Celtic background, MacNeice’s own literary influences and the poet’s influence on later authors, and the nature and the function of refrains and other forms of repetition. In analyses of MacNeice’s language, critics have emphasised the use of slant rhyme, alliteration, assonance and Celtic devices such as deibhide and aicill rhyme. Corcoran and Gillis have also commented on the poet’s systematic repetition of words, parts of words or of specific groups of words.

6.2.2. Basic annotations

The text corpus was created by scanning the eleven volumes of poetry that were selected. The case study focuses on *Poems*, *The Earth Compels*, *Autumn Journal*, *Plant and Phantom*, *Solstices*, *Holes in the Sky*, *Autumn Sequel*, *Ten Burnt Offerings Visitations*, *Springboard* and *The Burning Perch*. Machine-readable versions of the poems were obtained through Optical Characters Recognition (OCR). The results of the OCR were proofread to ensure that the electronic texts were free of scanning errors. In total, the corpus consisted of 311 poems, which

[402] Alan Gillis, ““Any Dark Saying”: Louis MacNeice in the Nineteen Fifties”, p. 113.
[403] Robin Skelton, “Celt and Classicist: The Versecraft of Louis MacNeice”.
[404] Ibid., p. 43.
[405] Ibid., pp. 44-45.

collectively contain 144269 tokens and 16623 types. This data set is not extensive, if measured against the criteria mentioned in Doug Laney's article on big data collections.[406] The relatively modest size of the corpus has had the advantage that the results produced by text analysis tools could be evaluated effectively by inspecting the original texts. A basic form of TEI encoding was also added to the plain texts. A program was developed in the PERL programming language to encode the texts strings that were delineated by hard returns in the original text file as verse lines, using the <l> element. All lines were also numbered. Each line received an identifier, which was captured in the @n attribute of the <l> element. POS tags and lemmas were added using the Morphadorner application, which was developed as part of the MONK project.[407]

A method was also developed for the creation of phonetic transcriptions of all the verse lines. The PoetryAnalyzer tool, which was developed by David Kaplan, could not be used, as it was based on a dictionary of American pronunciation.[408] A similar tool was developed by making use of the pronunciation dictionary that was developed for the MRC Psycholinguistic Database, which is available in its entirety from the Oxford Text Archive.[409] The format of the transcriptions from this dictionary was converted to SAMPA, which is a phonetic script based on the International Phonetic Alphabet, which makes use of ASCII characters only. Since it was found that the types that occurred in the corpus were not all available in the selected dictionary, a number of algorithms have also been implemented for the creation of phonetic transcriptions for the remaining tokens.[410] These algorithms, unfortunately, did not function perfectly, mainly because of the fact that there is no strong connection between spelling and pronunciation in the English language.[411]

[406] As one of the criteria, Laney states that data collections can be considered big if they are too voluminous to be managed on a single computer of average capacities. This was clearly not the case for the data produced in this study, as the total size does not exceed 50 MB. See Douglas Laney, *3D Data Management: Controlling Data Volume, Velocity, and Variety*, (2001). As the data collections that humanities scholars work with generally have a modest volume and a low velocity, it may be argued that the term "big data", as conceptualised within computer science and data science, is not fully applicable within the humanities. As an alternative to the term, Allen Riddell has coined the phrase "very large collection" to denote a collection which "contains more texts than a single researcher would be expected to digest in a year's worth of dedicated reading". See Allen Riddell, "How to Read 22,198 Journal Articles: Studing the History of German Studies with Topic Models", in: Matt Erlin (ed.), *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, Boydell & Brewer Ltd 2014, p. 92.

[407] <http://morphadorner.northwestern.edu> (16 February 2014)

[408] David Maxwell Kaplan, *Computational Analysis and Visualized Comparison of Style in American Poetry*.

[409] M.D. Wilson, "The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2", in: *Behavioural Research Methods, Instruments and Computers*, 20:1 (1988). The data set can be obtained at <http://ota.ox.ac.uk/desc/1054> (10 February 2013).

[410] The algorithms were based on the pronunciation guidelines that were provided in Beverley Collins & Inger Mees, *The Phonetics of English and Dutch* (Leiden: Brill 1998).

[411] Edward Carney, *A Survey of English Spelling* (Routledge 1994).

Mistakes in the transcriptions that were generated algorithmically were corrected manually.

The fragment below given an impression of how the basic linguistic annotations have been captured in TEI. The example is first line of the poem “MeetingPoint”.

```
<l n=“1”>
<w phon=“!taIm” lemma=“time” pos=“NNP”>Time</w>
<w phon=“wQz” lemma=“be” pos=“VBD”>was</w>
<w phon=“@-!weI” lemma=“away” pos=“RB”>away</w>
<w phon=“@n” lemma=“and” pos=“CC”>and</w>
<w phon=“!sVm-we@” lemma=“somewhere” pos=“RB”>somewhere</w>
<w phon=“!els” lemma=“else” pos=“RB”>else</w>
</l>
```

The data that were produced in this way provided the basis for additional forms of processing, which focussed on the detection of specific literary devices. The methods that were developed are discussed in more detail in the following section.

6.2.3. Perfect rhyme, slant rhyme and semi-rhyme

Many of MacNeice’s poems contain highly sophisticated experiments with poetic conventions and with the artistic effects that can be achieved through skilful combinations of similarities in sounds. Via computation, a number of the aural structures can be explored in a structural manner. Firstly, I have developed an algorithm for the identification of occurrences of perfect end rhyme. David Kaplan’s PoetryAnalyzer software detects cases of perfect rhyme by firstly extracting the final stressed syllable of the last word of each line, and by removing all consonants at the start of this syllable. In the case of feminine rhyme, this final stressed syllable is also followed by additional unstressed syllables. If specific phoneme sequences created in this way occur more than once, this is considered to be an instance of perfect rhyme. A similar logic was implemented in the application for this thesis’ case study. A number of modifications were found necessary, however. In Kaplan’s software, the algorithm operates on a window of four lines. Since it was found that MacNeice often uses rhyming schemes that span more than four lines, it was decided to extend the size of the window and to examine repetitions of rhyming schemes in full length stanzas. As rhyming schemes are generally defining at the level of stanzas, rhymes across different stanzas were disregarded. MacNeice has also written various poems which essentially consist of one lengthy stanza, such as “Valediction”. Such long poems have been divided, somewhat arbitrarily, into clusters of 8 lines. I assumed that rhymes in a larger number of lines were not close enough to be heard together. Using the procedure that was discussed, perfect rhyme was found in the words “implications” “revelations” in “A Contact”, “scales” and “tails” in “Sunday Morning”, “wells” and “cells”, “cease” and “caprice” in “April Manifesto”, and “halt” and “salt” in “Carrickfergus”. The procedure that was implemented also returned the rhyming scheme for each set of lines.

Further tests exposed a number of additional complications. The algorithm initially overlooked a number of lines which appeared to contain obvious examples of perfect rhyme. The poem “The Return”, for instance, contains lines ending in the nouns “consummations” and “patience”. The software application transcribed the former word as /*kɒn-sə-‘məɪ-ʃənz/ and the latter word as /’pəɪ-ʃəns/. Because of the different final phonemes, these line endings were considered to be sonically distinct. The final phoneme sequences sound very similar, however, and it seems incorrect to label this particular example as an instance of slant rhyme. It was decided to implement a somewhat less rigid form of matching, and to represent all sibilant sounds as a single /s/. Similarly, the diphthongs /əʊ/ and /ɪə/ were replaced, respectively, with /ɔ:/ and /i:/. Using this more lenient form of matching, a number of additional instances of perfect rhyme were found. Examples include the lines ending in “sprawls” and “Gauls” in “Museums”, “listen” and “vision” in “Visitations” and “choose” and “booze” in “Alcohol”.

Examinations of the results also revealed that MacNeice frequently uses repetitions of the exact same words at the end of poetic lines. It was decided that, if different lines contain identical words, this does not constitute perfect rhyme. In a number of cases, such repetitions may be considered instances of rime riche, as they consist of homonyms in which the spelling and sound are fully alike, while there is a significant difference in the meaning of these words.[412] In “Eclogue by a Five-Barred Gate”, for example, the poet writes that “My dream will word well / But not wear well / No dreams wear at all as dreams / Water appears tower only while in well” (ll. 85-87).

A similar method was implemented for the recognition of half rhyme, which, as was noted by Skelton, is an important and recurrent form of rhyme in MacNeice’s verse. The *Oxford Dictionary of Literary Terms* defines half rhyme, or slant rhyme, as a form of rhyme in which there is an agreement in the final consonants of the final stressed syllables, and a difference in the vowel sounds. An alternative form of slant rhyme is one in which there is a match in the vowel of the stressed syllable, and a difference in the consonants. In this study, I refer to the first form as consonance rhyme, and to the latter form as assonance rhyme. To find lines that contain repeated consonants, a method was developed which firstly selects the final stressed syllable. The initial consonants of this syllable are retained. In this syllable, all vowel sounds are replaced with a placeholder character, and the resulting pattern is used to find matches. To detect assonance rhyme, a text pattern is created in which all consonants are supplanted by placeholder characters. Obvi-

[412] *The Princeton Encyclopedia of Poetry and Poetics*, p. 1199.

ously, all instances of perfect rhyme needed to be removed before the start of the procedure.

Instances of assonance slant rhyme were found in the words “return” and “worth” in “Ode”, in “clover”, “sober” and “cream-soda” in “The Individualist Speaks”, and in “wisps” and “fist” in “Train to Dublin”. The algorithm also showed correctly that the well-known poem “Bagpipe Music” makes an almost exclusive use of feminine assonance rhyme (“python” and “bison”, “sofa” and “poker”, “whiskey” and “fifty”, “Blavatsky” and “taxi”, “Ceilidh” and “baby”). The use of slant rhyme effectively underlines the centre theme of the poem, as it focuses on the decline of a traditional rural culture of the Western islands of Scotland, and the hasty transition to a consumerist and urban culture. In this poem, the slant rhymes evoke a confused disorganisation, and suggest an abandonment of order and predictability. The insistent vowel sounds also mirror the lingering drone of the bagpipe. Consonance rhyme was detected in, among many other poems, “Birmingham”: “On shining lines the trams like vast sarcophagi move / Into the sky, plum after sunset, merging to duck's egg, barred with mauve” (ll. 25-26). It was also found in words such as “hourly”, “rarely” and “merely” in the poem “Entirely”and in “burden” and “garden” in “Prognosis”.

Instances of semi-rhyme were found by comparing the line endings that contain a single syllable to line endings that contain multiple syllables. When specific masculine rhymes match the stressed syllables in a feminine rhyme, this agreement is considered to be an example of semi-rhyme. Using this method, this form of rhyme was detected in lines that end in the words “eye” and “island” (“Street Scene”, ll. 32-33), “sky” and “horizon” (“Ode”, ll. 50-52), “way” and “neighbours” (“Eclogue from Iceland”, ll. 220-221), “skin” and “windows” (“Country Week-End”, ll. 84-86) and “run” and “Sunday” (“Sunday in the Park”, ll. 13-15). Evidently, this algorithm also returns line endings in which the stressed syllables are precisely the same, as in the final lines of the poem Mayflies (“May” and “mayflies”). It was decided, nevertheless, to count such cases as forms of semi-rhyme as well. The device is used frequently to emphasise the simultaneous sameness and diversity of phenomena. The poem “Wolves” contains a clear example. In this poem, the persona finds “pathos in dogs and undeveloped handwriting / And young girls doing their hair and all the castles of sand”. In this example, the matching acoustic patterns exemplify the inherent correspondences between the various phenomena which may shield the poet against the relentless unpredictabilities of reality. The poem “Snow Man” contains an equally functional use of semi-rhyme. The poem reflects on the nature of human memory and on the changing image of the self, by comparing it to a melting snow man. The second stanza muses on the decay of the remembered self-image: “Tomorrow / Comes the complete forgetting, the thaw / Or is it rather a dance of water”. In these lines, the solid shape of the single syllable dissolves, like melting snow, into a more fluid and more ephemeral state. Systematic manipulations of phonetic transcriptions can clearly help to demonstrate the verbal skills of MacNeice and his ability to create engaging acoustic effects.

6.2.4. Deibhide rhyme and internal rhyme

Robin Skelton has argued that MacNeice's Irish background manifested itself strongly in his frequent use of Celtic devices such as deibhide rhyme and aicill rhyme. Deibhide, firstly, is a form of rhyme in which a stressed syllable rhymes with an unstressed syllable. To trace occurrences of this first type of rhyme, an algorithm was developed which initially selected the final syllables of all lines within a stanza, regardless of whether these were stressed or unstressed. All initial consonants were removed. Lines were taken to contain deibhide rhyme if there was a match in the phoneme sequences of stressed and unstressed syllables, as in the following cases:

Holidays should be like this / Free from over-emphasis ("PostScript from Iceland", ll. 37-38)
The trains pass and the trains pass, chains of lighted windows (...) / For these are the trains in which one never goes ("A Contact", ll. 1-3).
But they all go so fast, bus after bus, day after day / Year after year, that you cannot mark any headway ("The Glacier", ll. 7-8)

During a review of the initial results, it was found that the algorithm disregarded a number of cases which seemed to contain obvious occurrences of deibhide rhyme. The poem "No More Sea", for instance, contains lines ending in "sum" and "medium". The final syllable of "medium" is transcribed phonetically using an unstressed schwa, and for this reason, it does not match the vowel in the word "sum". Problems such as these are difficult to solve. One option would be to work with a highly forgiving form of matching in which the schwa and the /ʌ/ vowel are viewed as identical, but this would result in matches on other locations which are clearly not intended as rhymes. To recognise noteworthy sonic patterns, the computer must be lenient in some cases, and strict in other situations. Capturing the appropriate level of flexibility clearly remains elusive.

Skelton also argues that the influence of the Celtic verse tradition is noticeable though the use of aicill rhyme. The *Princeton Encyclopaedia of Poetry and Poetics* explains that aicill rhyme consists of a correspondence between a phoneme sequence at the end on one line and a word in the interior of the line that follows immediately.[413] To study instances of this device, I examined windows of two lines. From the first line, the final stressed syllable was taken, which could optionally be followed by unstressed syllables. This pattern was compared to phoneme sequences in all the unique words of the following line. The last word of the second line was obviously neglected, as an agreement in the final words of two subsequent lines would constitute a regular perfect rhyme. It was also decided to disregard

[413] *The Princeton Encyclopedia of Poetry and Poetics*, p. 26.

lines that end in a single unstressed syllable. Among many other examples, the following instances could be detected:

Far more fatal than painted flesh or the lodestone of live hair / This despair of crystal brilliance ("Circe", ll. 4-5)
Filleted sun streaks the purple mist / Everything is kissed and reticulated with sun ("Morning Sun", ll. 8-9).
Stressing the function scrapping the Form in Itself / Taking the horse from the shelf and letting it gallop (Autumn Journal, section XII, ll. 50-51)

Internal rhyme, or occurrences of perfect rhyme within a single line, may be found by firstly compiling a list of all the unique words within a line, and by finding the phoneme sequences that occur more than once. This form of internal rhyme was found in the following fragments:

Its cage is a stage its perks are props ("Budgie", l. 4)
And the sun stood still above Notting Hill Gate ("Jericho", l. 26)
In quiet in diet in riot in dreams ("The Creditor", l. 4)

Skelton also discusses a specific form of aicill rhyme in which the final consonant sequence of one line reoccurs in the lines that follow. An attempt was also made to find this particular form of repetition. I implemented a procedure which analyses couplets of consecutive lines. All vowels are removed from the final words of the first line and from the entire second line. If the sequence of consonants is found to reoccur in the second line, this is captured as a case of consonantal aicill rhyme. During the initial tests, it became clear that there were many cases in which the last word of the first line and the first word of the second line were identical. Following Eagleton, who insists that rhyme entails "a unity of identity and difference",[414] it was decided to ignore all cases in which words are repeated verbatim and to focus solely on sonic affinities in distinct words. Next to the examples which were also discussed by Skelton, this form of aicill rhyme was found in the lines "Of the light in the dark of the muted voice of the turning wild / World yet calm in her storm gay in her ancient rocks" ("All Over Again", ll. 17-18) and "Your smile and chivvy your limbs through a maze of pearly / Pillars of ocean death--and yet you force your way" ("The Casualty", ll. 29-30)

6.2.5. Metre

In this case study, I also attempted to automate parts of the scansion. The stress patterns of the verse lines could be obtained directly from the phonetic transcrip-

[414] Terry Eagleton, *How to Read a Poem*, p. 132.

tions that were produced in this study, since these included data on the primary and the secondary stresses in each word. The stress patterns that were found were also classified. Methods were developed for the recognition of iambs, trochees, dactyls and anapests in trimeters, tetrameters, pentameters, hexameters and heptameters. Both regular verse lines and their catalectic or incomplete versions were considered. As I found that the algorithm initially resulted in very few examples of perfect iambic pentameters, I implemented a more lenient method, in which one of the feet in the pentameter was allowed to contain one supplementary unstressed syllable. Using this logic, iambic pentameters were found in the lines "To every question gives the same reply" in line 6 of "The Conscript" or "You cannot argue with the eyes or voice" in line 289 of "Eclogue from Iceland". The line "We are dying, Egypt, dying" ("The Sunlight on the Garden", l. 18) was found to be an example of a catalectic iambic trimeter, and "Neither sense nor conscience stirred" ("Jericho", l. 33") was classified as catalectic trochaic trimeter.

This method for the recognition of the metrical character of the various lines did not function flawlessly. An obvious difficulty is that the regular stress patterns of words may change within the context of a particular verse line. The poem The Sunlight on the Garden, for instance, is largely written in catalectic trimeters. The word "and", in the poem's second line, "hardens and grows cold", clearly receives stress within this rhythm, while it is listed as an unstressed word in the pronunciation dictionary. Complications such as these are difficult to solve, however, without a manual rearrangement of the stress patterns in the phonetic transcriptions.

6.2.6. Alliteration, assonance and consonance

In his essay "Feeling into Words, Nobel laureate Seamus Heaney has noted that the Ulster accent is "generally a staccato consonantal one".[415] Devices such as alliteration and consonance are correspondingly highly common among poems from Northern Ireland. In this study, I also attempted to analyse the ingenious repetitions of constants and of vowels within MacNeice's poems. Occurrences of alliteration were found, firstly, by counting the frequencies of all the consonants and the vowels that occur at the beginning of syllables that have primary or secondary stress. In this study, it was decided that when words are repeated, these repeated sounds do not constitute alliteration. For this reason, all repeated words were removed. I also assumed that alliteration takes place when sounds at the beginning of stressed syllables occur more than once. Following this logic, a heavy use of alliteration was found, for instance, in "Prayer Before Birth" ("With Strong drugs dope me with wise lies lure me, on black racks rack me in blood-baths roll me" / [...] / With water to dandle me grass to grow for me trees to talk" (ll. 6-9).

[415] Seamus Heaney, "Feeling into Words", in: *Finders Keepers*, London: Faber and Faber 2003.

Alliteration was also found in the second line of the poem Belfast: “Frozen into his blood from the fire in his basalt”. This line is particularly interesting as the sounds that are repeated are the same as the main sounds in the name of the city that is depicted in the poem. Abrams suggests that the term ‘alliteration’ mainly refers to repetitions of consonants. Restricting the algorithm to occurrences of consonants, however, would result in inattentiveness to a number of striking instances, such as “Anger and ambush” in line 18 of “Iceland”, “Who was innocent and integral once”, in line 137 of “Ode” and “He plods the endless aisles not daring to close and eye” in line 2 of “Flower Show”. It was decided, for this reason, to search for matches both in consonants and in vowels.

The method that was implemented in this study returned a pattern that represented the specific sounds that were repeated. On the basis of the complete list of patterns, it can be concluded that the poet deliberately planned alliterations. MacNeice frequently applied patterns in which particular pairs of sounds are repeated, or in which two alliterative words are nested within two other words whose initial consonants also alliterate. For lines 20 and 21 of “Western Landscape”, for instance, (“And hanging smell of sweetest hay / Weavingly laughingly leavingly weepingly”) the software produced the pattern “- h s - s h / w l l w”. Examples of repetitions of pairs of sounds can be found in the lines “Prowl and plunge through glooms and gleams” (l. 20), and “A halfway house between sky and sea being of the water earthy” (l. 80) in “Donegal Triptych”. Instances of alliterations nested within other alliterations can be found in the lines “Posed by Picasso beside an endless opaque sea”, in “An Eclogue for Christmas”, l. 33 or “In certain long low islets snouting towards the west”, in “Last before America”, l. 19. A similar, less noticeable, structure occurs in “Let us too make our time elastic” in “Mayfly”, l. 23.

Similar methods were developed for the recognition of assonance and consonance. The *Princeton Encyclopedia* defines consonance as “the repetition of the sound of a final consonant or consonant cluster in stressed, unrhymed syllables near enough to be heard together”.[416] Following this definition, I developed an algorithm which considers all possible windows of four consecutive stressed syllables in each verse line. I decided that the lines contains consonance or assonance if at least one of such clusters contains a repetition of a consonant or vowel at the end of these syllables. Using this logic, assonance was found in the lines “this dyspeptic age of ingrown cynics” (“Eclogue from Iceland”, l. 117) and “In the sun-peppered meadow the shepherds are old” (“Nuts in May”, l. 5). Similarly, consonance was found in the lines “Have seen myself sifted and splintered in broken facets” (“An Eclogue For Christmas”, l. 34), “Metal patents parchment lampshades harsh” (“Belfast”, l. 11), “The tight-lipped technocratic Conquistadores” (“Epitaph for Liberal Poets”, l. 13).

[416] *The Princeton Encyclopedia of Poetry and Poetics*, p. 299.

6.2.7. Repetitions of words

Neil Corcoran has discussed MacNeice's use of refrains. It is a device in which a line, or a part of a line, is repeated verbatim within a poem.[417] Corcoran surmises that "repetition at the formal or technical level can be thematically functional",[418] and illustrates this statement via discussions of the poems "Leaving Barra", "Train to Dublin", "The Hebrides" and "Meeting Point". As part of the experimentation that was conducted for this thesis, I devised a method to identify repeated occurrences of n-grams of two or more words. Since refrains are taken to consist of literal repetitions, all function words, such as articles and pronouns, were retained during the creation of n-grams.

Corcoran's article T*he Same Again? Repetition and Refrain in Louis MacNeice* discusses 11 poems that contain refrains. The software that was developed for the case study was able to recognise the refrains in these same poems, and also identified a large number of additional examples. In the poem "April Manifesto", the word sequence "our april must replenish" occurs three times. The poem uses imagery of spring and of regeneration to represent the aesthetic delight that may be derived from an experience of colours and of sounds, and combines this with a mild criticism of consumerism. The repetition of the refrain at the start, middle and end of the poem effectively stresses the insistence of the craving for abundance. "An Eclogue for Christmas" contains several repetitions of the phrase "What will happen", and this underscores the poem's central sense of doom and anxiety about the future. In "Evening in Connecticut", the line "only the shadows growing longer and longer" is used both in the first and in the final stanza. At the beginning of the poem, these words are used mainly to describe the tranquillity of the evening and of the natural surroundings. The poem as a whole, however, mainly reflects on the immanence of a World War in Europe, and towards the end of the poem, the lengthening shadows essentially symbolise the darker qualities of human nature.

In some cases, the repetitions that were recognised by the application did not always constitute clear examples of refrains. In the poem "Invocation", for example, repetitions of the phrase "fetch me far" were found at the beginning of 14 lines in the poem. These repetitions, which collectively underscore the dream-like and escapist surge of the poem, ought to be classified more appropriately as anaphora. The same can be claimed for the repetitions of "It's no go" in "Bagpipe Music". In this poem, according to Neil Corcoran, the "almost demented repetitiveness propels the poem's hurdy-gurdy rhythmic relentlessness, as if the repeated phrase has taken over the poem".[419] If it was found that two or more lines contain the exact same words at the beginning of the line, such cases were captured as

[417] *The Princeton Encyclopedia of Poetry and Poetics*, p. 1151.

[418] Neil Corcoran, "The Same Again? Repetition and Refrain in Louis MacNeice", p. 216.

[419] Ibid.

instances as anaphora. Other examples of anaphora were found in “June Thunder” (“If only you would come [...] / If only now you would come”, ll. 24-25), in “Plurality” (“Conscious of guilt and vast inadequacy [..] / Conscious of waste of labour”, ll. 73-75) and in “Twelfth Night” (“O crunch of bull’s-eyes in the mouth / O crunch of frost beneath the foot”, ll. 2-3).

The software also pointed to an echo of “their verdure dare not show” at the beginning of “Valediction”, but this phrase is repeated within a single verse line. Since it is stressed in the *Princeton Encyclopaedia of Poetry and Poetics* that refrains ought to be separated “by at least one line of nonrepeating material”,[420] the application was enhanced with measures to ensure that the refrains do not occur on the exact same line. Furthermore, the simple fact that sequences of words are used more than once does not automatically imply that these are also meaningful refrains. The software identified recurrences of the words “here it was” in “A Hand of Snapshots”, and of the phrase “over the wall” in The Stygian Banks. The reasons why these particular phrases should not be viewed as refrains, however, are difficult to codify in an algorithm. This finding stresses the continued need for manual evaluation and correction of the results.

In his discussion of MacNeice’s “Leaving Barra”, Corcoran also notes that the poem contains a specific form of repetition in which words or parts of words are repeated in other words within the same line. Examples can be found in the phrases “the rain and the rainbow” and “a belief that is unbelieving”. Gillis argues that, in MacNeice’s later writing, such forms of repetition “becomes a pivotal means of exploring emptiness and destabilization in the late Fifties”. While Corcoran does not propose a term for this stylistic device, it may expediently be referred to as paronymy. This term refers to “two or more words partly identical in form and/or meaning, which may cause confusion in reception or production”.[421] The recognition of this stylistic device can partly be automated. A first version of the algorithm simply established whether or not a word was contained in any of the other unique words on the same line. This method, however, was too crude, as it was found that words that consist of a single character, such as the article ‘a’, and the personal pronoun ‘I’, are obviously contained very frequently within other words. Since it was assumed that the device predominantly occurs in nouns, verbs and adjectives, a second version of the algorithm made use of the POS tagging, and initially removed prepositions, conjunctions and articles from all lines. This second method also overlooked a number of cases which did seem relevant, however. The lines “Beyond these plains’ beyondless margin” and “Yet standing here and notwithstanding / Our severance” in “Letter from India”, or “So much themselves in despite of spite” in “Visitations” were ignored, because of the removal of prepositions. A clear difficulty is that prepositions are relevant in some cases, but

[420] *The Princeton Encyclopedia of Poetry and Poetics*, p. 1151.

[421] R. R. K. Hartmann & Gregory James, *Dictionary of Lexicography* (Routledge 2002), p. 192.

irrelevant in other situations. The repetition of “in” on line 13 of “June Thunder”, for instance, (“Then the curtains in my room blow suddenly inward”) does not seem of significance.

An improved version of the algorithm made use of a list of stopwords. The standardised Glasgow list of stopwords, which is also employed in the Voyant software, proved to be too inclusive for this particular purpose.[422] Since words such as “sometime”, “part”, “together” and “give” are all on the Glasgow list, using this resource had the effect that striking formulations, such as “Next year is this year, sometime is next time, never is sometime” in line 76 of “Homage to Clichés”, “Made him a part of the not to be parted whole” in line 71 of “Western Landscape”, “Forgive what I give you” in line 1 of “To Mary”, and “The more there are together, Togetherness recedes” in line 7 of “Babel” were all ignored. It was decided, for this reason, to make use of an edited list of stop words.

An important shortcoming was that the methods that have been discussed all failed to recognise one of the examples which was highlighted by Corcoran. The phrase “a belief that is disbelieving” is not recognised as paronymy, as “belief” is obviously not repeated in its entirety in “disbelieving”. Making use of lemmatisation also would be ineffective in this particular case, as the root form of the latter word is “disbelieve”. As a solution to this difficulty, I developed a method in which all possible substrings were extracted from words that contained more than three characters, excluding words on the edited list of stopwords. These substrings were used as the basis for the comparison. This procedure retrieved a number of noteworthy lines from the corpus, including “Of all desirable things - that is what I desire” in “Troll’s Courtship”, “Our past we know / But not its meaning — whether it meant well” and “Memories I had shelved peer at me from the shelf” in “Carrick Revisited”, “what to these does the word significant signify” in “The Stygian Bank”, “All the unconsummated consummations” in “The Return”, “More than the twanging dazzle or the dazzling noise” in “Ode”, and “the scalloped / Lampshade swings a shadow” in “Trilogy for X”.

6.2.8. Onomatopoeia

Algorithms may similarly be created for the detection of onomatopoeia, or words whose phonetic aspects mimic the sound of the things they refer to.[423] The sounds of the words when pronounced could obviously be derived from the phonetic transcriptions, but data on the sounds produced by the referents of words were clearly unobtainable. There appear to be two distinct cases of onomatopoeia.

[422] Stop lists are resources which list the words which are most frequent within a particular language, and which, according to its developers, are of less importance for an analysis of the semantic contents of texts. The Glasgow list of stopwords was developed by The Information Retrieval Group at the University of Glasgow and contains 319 words.

[423] “Onomatopoeia”, in Chris Baldick, *The Oxford Dictionary of Literary Terms*.

Firstly, specific words are onomatopoeic regardless of their context. Secondly, there are also words which assume onomatopoeic qualities only because of their use within a specific context. One example of the latter type of onomatopoeia can be found in “The Cyclist”, which narrates a scene in which a boy cycles down a hill past the Westbury White Horse. The first line of the poem, “Freewheeling down the escarpment”, effectively evokes the wind racing past the cyclist. While it seems virtually impossible to detect the latter type of onomatopoeia, the first type may potentially be detected by making use of a terminable list of words whose sounds resemble the acoustic aspects of the thing they represent. The list that was used for the purpose of the current case study also includes all inflected forms of verbs and both the singular and verbal forms of nouns, in order to ease processing. Instances of onomatopoeia were found in phrases such as “crickets fiddled and sizzled to drown the river” (“The Rest House”, l. 5), “the squelch of mud the belch of surf” (“A Hand of Snapshots”, l. 45) and “semaphore ultimatums tick by tick” (“The Hebrides”, l. 108). Polysemous words such as “flush” or “spark” are clearly not onomatopoeic in all contexts. The lines “No spark of reality possible” (“Eclogue by a Five Barred Gate”, l. 62) or “To keep it flush with the earth” (“Under the Mountain”, l. 6) contain words which were included in the list of onomatopoeic words, but do not form compelling examples of onomatopoeia. As was the case for other figures of speech, the list of results that was produced by the software still needed to be examined and amended manually.

6.2.9. Allusions

During his lifetime, MacNeice produced an impressive body of literary criticism. *The Selected Literary Criticism of Louis MacNeice*, edited by Alan Heuser, includes essays on W.H. Auden, T.S. Eliot and Dylan Thomas, and following the death of W.B. Yeats in 1939, MacNeice also wrote an extensive critical study of the poetic works of his compatriot. It may be assumed that MacNeice’s voracious reading also helped to shape the nature of his poetic development. Several authors have noted, for instance, that MacNeice’s book on Yeats reveals as much about the older poet as about the author himself.

Among many other ways, MacNeice’s literary influences may be explored by studying the explicit allusions to other literary texts. Crane argued that algorithmic explorations of allusions can be based on lexical similarity, on syntactical similarity or on phonetic agreements. In this case study, the attempt to identify allusions was based on agreements in vocabulary. An experiment was conducted which concentrated on the detection of the lexical parallels between poems by MacNeice and by W.B. Yeats. The algorithm that was developed firstly identified all the lines in the works of Yeats and of MacNeice which share two or more words. If there are clear lexical parallels in the usage of words, such correspondences may, in some cases, be characterised as allusions. To attenuate the impact of inflections, the algorithm made use of the lemmatised versions of the verse lines. In addition, all stop words were removed, using the same manually edited list that was used for the detection of refrains. It was found, however, that this initial method resulted in an inordinately high number of combinations. To reduce the number of results, I added a requirement which stipulated that the shared lemmas should also be used in the same sequence. The number of verse lines that share such sequences of two lemmas was still decidedly high. Within the 16,782 lines by MacNeice and the 11,937 lines written by Yeats, I found 8,633 instances of lines with shared words. No matches of three or more words were found, however.

The algorithm assuredly helped to identify a number of compelling resemblances between poems by MacNeice and by Yeats. The opening line of the late poem "Flower Show", for instance, ("Marooned by night in a canvas cathedral under bare bulbs") contains an ironic invocation of the sixth part of Yeats's "Under Ben Bulben", in which the poet describes his epitaph ("Under bare Ben Bulben's head / In Drumcliff churchyard Yeats is laid", l. 91). The method that was developed confirmed Edna Longley's observation that the line "All you do is burke the other and terrible beauty" in "Eclogue by a Five Barred Gate" alludes to Yeats's line "A terrible beauty is born" from "Easter 1916". [424] The results produced by the algorithm also showed, interestingly, that the line that follows ("all you do / Is shear your sheep to stop your ears") (ll. 37-39) shares two consecutive lemmas with Yeats's "To a Shade". The latter poem consists of a reprimand against the Irish people who scorned Parnell, the founder of Irish Parliamentary Party. The poet advises the ghost of Parnell to "gather the Glasnevin coverlet / About your head till the dust stops your ear" (ll. 21-22). While "To a Shade" urges the ghost of Parnell to escape contemporary political realities, MacNeice's eclogue is conversely an exhortation to the shepherds to become more politically involved.

Many additional verbal parallels were found, some of which were significant. Common words were found, for instance, in "The Closing Album" and in "The Nineteenth Century and After". Part V of the former poem depicts the sentiment which is also expressed in the opening chapter of MacNeice's book on Yeats. MacNeice wrote that, after he had heard the news about the outbreak of the Second World War, this news established a new reality, which made nonsense of the old reality. "The Closing Albums" asks why the sea must continue to "draw a film of muslin down the sand / With each receding wave?". A parallel was found in Yeats's brief poem "The Nineteenth Century and After", which similarly focuses on the end of a period. Yeats urges readers, nevertheless, to appreciate the present and to value the "rattle of pebbles on the shore / Under the receding wave". Similarly, the line "That Man is a dancer is an anachronism" in "Precursors", may be viewed as an allusion to Yeats's "Nineteen Hundred and Nineteen", which asserts that "All men are dancers". The phrase in the Yeats poem is used in connection with the "Platonic Year", which, according to Jeffares, represents the notion that "the whole of the

[424] Edna Longley, *Louis MacNeice: A Study* (London: Faber and Faber 1988), p. 102.

constellation returns to the positions from which they once began".[425] MacNeice's poem suggests that such cultural and ideological renewal can no longer be achieved in the late twentieth century. The "topless tower" in "Brother Fire" may be viewed as a reference to the "topless towers / Where Helen walked with her boy" which are depicted in Yeats's "When Helen Lived". Both poems focus on a destructive force within human nature, which obstruct a full commitment to beauty in times of hardship. Section XIV of *Autumn Journal* depicts a journey through rural Oxfordshire, and on line 11, the car's windscreen wiper is compared to a "cricket that sings". This image clearly echoes the line "Dropping from the veils of morning to where the cricket sings" in "The Lake Isle of Innisfree". Both poems, notably, concentrate on the redemption that may emanate from a retreat into nature. "Old Masters Abroad" likewise contains a references to "The Lake Isle of Innisfree" in the line "Nine bean rows rise in the Kalahari".

As was shown, a method which connects lines that share two or more consecutive lemmas can disclose noteworthy similarities between the works of different authors. Whereas such lexical parallels do not always indicate intended allusions, placing such different literary contexts side by side can occasionally quicken reflection on similarities or contrasts. A first problem with the method that was followed, however, is that it returns a high number of shared bigrams, while only a minority of these cases seem significant from a critical point of view. Unfortunately, as noted by Coffee et al., the precise quality that renders an allusion relevant or significant often remains incommensurable.[426] The phrase "We are dying, Egypt dying" in "The Sunlight on the Garden" is an obvious allusion to Shakespeare's "Anthony and Cleopatra", but the algorithm that was implemented in this experiment also singled out lines containing common collocations such as "some day", "close eye", "so long", "young man", "come from", "each other", "all day" and "far away". Moreover, intertextual reference are frequently highly complex, and a method crudely based on shared bigrams callously misses such more intricate types of allusion. The opening line of the second section of "Autumn Journal", "Spider, spider, twisting tight" is an obvious allusion to William Blake's "Tiger", but a method based on verbal similarities would evidently fail to link the two lines. The first stanza of "Brother Fire" depicts the fires during the London Blitz as a dog raging through the streets, and this imagery is reminiscent of the phrase "let slip the dogs of war" from Shakespeare's "Julius Caesar". The recognition of such extended allusions clearly demands a more sophisticated matching algorithm. Richard Danson Brown also notes that the third stanza of "Neutrality" echoes "the vocabulary and idiom" of Yeats and that the phrase "Intricacies of gloom and glint" is reminiscent of the line "In all lovely intricacies of a house" from

[425] W.B. Yeats, Norman Jeffares (ed.), *Yeats's Poems* (Basingstoke: Macmillan 1996), p. 583.

[426] Neil Coffee et al., "Modelling the Interpretation of Literary Allusion with Machine Learning Techniques".

Yeats's "In Memory of Major Robert Gregory".[427] Brown suggests that allusion may also be based on a singular common word. The precise reasons why repeated words constitute allusion in some cases, while they do not in other situations, seem impervious to formalisation. A method based on shared n-grams may help to identify a number of striking parallels, but the task to separate the significant matches from the insignificant matches is highly labour-intensive. In addition, the method also missed many more perplexing categories of intertextual reference. As an idea can be expressed in many different ways, using very different words, a method that is based on ngrams only has limited value. Because of these difficulties, the attempt to recognise literary allusions automatically has not been pursued further.

6.2.10. Imagery

In his monograph *Sceptical Vision*, Terence Brown devotes a full chapter to the discussion of MacNeice's imagery. One of the images that Brown concentrates on is that of the sea, which, as he argues, features in many poems as "a major image of eternity, of the beyond, of non-being".[428] In "Western Landscape" and "Around the Corner", conversely, the sea is mostly portrayed as a redeeming force.[429] Brown adds that other natural elements, such as wind and stones, frequently have a particular significance as well. Compelling examples of references to stones and to petrification can be found in "Nocturne", "The Glacier" and "Western Landscape". Brown also discusses the poet's references to various modes of transportation, such as cars, boats and trains. In the poem "Trilogy for X" the train represents the notion that the persona moves "through a world of vanishing particulars where new data or new phenomena present themselves continually".[430] A related image is that of the quest, which often represents the poet's uneasiness with an ever-changing reality and the central hankering after profundity.

Computation may enable scholars to trace references to specific images, but a computer-based analysis of imagery is complicated by a degree of opacity with respect to the precise definition of the term. Baldick notes that imagery is a "rather vague critical term covering those uses of language in a literary work that evoke sense-impressions by literal or figurative reference to perceptible or 'concrete' objects, scenes, actions, or states".[431] In his essay "Modern Poetry", MacNeice also makes an important distinction between, on the one hand, the properties of a poem, which are essentially the object and the actions which are needed to construct a narrative, and, on the other hand, the images proper, which are words

[427] Richard Danson Brown, "Neutrality and Commitment: MacNeice, Yeats, Ireland and the Second World War", in: *Journal of Modern Literature*, 28:3 (2005), p. 114.

[428] Terence Brown, *Louis MacNeice: Sceptical Vision*, p. 118.

[429] Ibid., p. 121.

[430] Ibid., p. 106.

[431] "Imagery", in Chris Baldick, *The Oxford Dictionary of Literary Terms*.

which primarily have a metaphorical function and which represent more recondite concepts.[432] As it seems unattainable to make a distinction between properties and images on formal grounds, however, these two types of images will be treated equally.

In this case study, a first attempt to extract imagery from the corpus was based on topic modelling. In this experiment, the MALLET program has been set to recognise 100 topics. An evaluation of the results of MALLET revealed that the words which occur frequently in the same documents were placed in one topic. None of the word clusters that were produced, however, could easily be resolved to a recognisable image. As was also demonstrated by other authors,[433] topic modelling and LDA do not seem helpful for the study of imagery in corpora containing figurative language.

In this study, attempts to identify imagery algorithmically were also based on the UCREL Semantic Analysis System (USAS) and on The Harvard General Inquirer (HGI). USAS, firstly, as was discussed in Chapter 3, is a semantic tagger which can connect the concrete tokens found in a document to broader semantic categories.[434] The entire MacNeice corpus was tagged using USAS, and categories that were applied most frequently included A3 ("Being"), B1 ("Anatomy and physiology"), M6 ("Location and direction"), M1 ("Moving, coming and going"), O2 ("Objects generally"), O4 ("Physical attributes"), T1.3 ("Time:Period"), L3 ("Plants").[435] Categories which refer to abstract concepts or to specific components of argumentative disposition, such as the categories in section A1 ("General and abstract terms") and Z ("Names and grammatical terms") were largely ignored, since this study is interested in words that that refer to concrete objects or events. The initial analysis offers some support for Brown's observation that MacNeice was preoccupied with movement, travelling and time. The USAS classifications also exposed references to types of imagery that was not discussed in the criticism that was surveyed for this study. It was found, for instance, that poems also contain many references to plants and to flowers: "The murderous grin of toothy flowers" ("Intimations of Mortality, l. 14), "Coral azalea and scarlet rhododendron" (Ode, l. 99), "There is more than glass between the snow and the huge roses" ("Snow", l. 12), "Frost will not touch the hedge of fuchsias" ("Valediction, l. 76), "a welter of nasturtium" (The Closing Album, l. 89). One difficulty with USAS is that the coverage of its categories is decidedly broad for this specific purpose. The USAS categories do not correspond directly to the imagery that is discussed by earlier critics of MacNeice's verse. Browns focuses on highly specific images, such as that of the sea, storms or church bells, but, within USAS, such phenomena or objects

[432] Louis MacNeice, *Modern Poetry: A Personal Essay*, p. 91.

[433] Lisa M. Rhody, "Topic Modeling and Figurative Language".

[434] A guest account to WMatric and to USAS was kindly provided by dr. Paul Rayson, Reader in Computer Science at the University of Lancaster.

[435] <http://ucrel.lancs.ac.uk/usas/semtags.txt> (18 July 2014))

are mostly subsumed within broader terms. Terms pertaining to quest imagery, for example, partly correspond to the lexicons for categories such as M1 (“Moving, coming and going”), M2 (“Putting, taking, pulling, pushing, transporting &c. “), M3 (“Movement/transportation: land”) and M4 (“Movement/transportation: water “), but these headings also cover many terms that are clearly unrelated to the notion of the quest.

Imagery may also be detected, to some extent, by making use of the HGI, which is “a computer-assisted approach for content analyses of textual data”.[436] The lexicons can be used to identify words with a negative or a positive connotation, words which express strength or weakness, or words which indicate a passive or an active orientation, among many other categories. A minority of lexicons contain terms that refer to more concrete phenomena or objects. One subclass enumerates words that describe “created locations that typically provide for social interaction and occupy limited space”.[437] It includes lists for tools, types of food, vehicles and buildings. Furthermore, the HGI also provides a list of words denoting colours. The tool can usefully be applied to recognise tokens in the general categories which were defined within HGI. As is the case for USAS, however, the HGI operates on semantic fields which are generally much broader than those which interest literary critics. One of the images which are discussed by Terence Brown, for instance, is that of the Madonna. The image figures prominently in poems such as “Belfast” and “Evening Indoors”. Within HGI, by contrast, the term “madonna” is subsumed under the much broader category “religion”.

In this study, a solution was developed in which a number of customised lexicons were developed, using the lexicons that were supplied by USAS and by the General Inquirer as a basis. All categories from the USAS system which do not refer to concrete events of objects, such as A13.1 (“Degree: Non-specific”) or N5 (“Quantities”) were removed. Lexicons from USAS and from the General Inquirer which refer to similar phenomena, such as “Religion” and “religion and the supernatural” were merged. Broad categories in the sections “M” and “T” were replaced manually with narrower categories, so that searches for specific images such water, wind, stones, threads also became possible. These bespoke lexicons, which cluster words that evoke specific objects, events or sensory impressions,[438] helped to identify many of the images which are used in MacNeice’s poetry.

To gauge the accuracy of this method, all poems which are discussed in the chapter “The Poet and his Imagery” from Brown’s monograph on MacNeice were listed, in combination with the images these contain. Additionally, an application was developed in which the results of the method based on machine reading were

[436] <http://www.wjh.harvard.edu/~inquirer> (21 July 2014)

[437] <http://www.wjh.harvard.edu/~inquirer/homecat.htm> (21 July 2014)

[438] The method that was used in the case study of this thesis is similar to the method that is described in John B. Smith, “Computer Criticism”, p. 20.

compared with the results of Brown's readings of the poems. This comparison showed that the software managed to identify the same images as Brown in the majority of cases. This was the case for poems which, according to Brown, contain references to threads, church bells, the sea, trains and boats. Additionally, the software pointed to additional references to these same images in many other poems. It identified 40 poems containing thread imagery which were not listed in Brown's study. Brown explains that the image of a thread or a ball of wool often represents "life in perpetual change in time"[439], but it is also used to describe the permanence of the ancestral background in "Valediction": "The woven figure cannot undo its thread" ("Valediction", l. 43). In "Postscript to Iceland", by contrast, the line "All the wires are cut, my friends" (l. 67) indicates a separation from past securities. The software for the detection of imagery also pointed towards 43 additional poems which confirmed Brown's observation that church bells often have a sinister connotation. It has this effect in "Sunday Morning" ("the church spire / Open its eight bells out skulls' mouths which will not tire", ll. 10-11), "Half Truth from Cape Town" ("Between a smoking fire and a tolling bell", l. 1) and section XVI of "Autumn Journal" ("yet her name keeps ringing like a bell", l. 59). The results of the algorithm were not equally convincing for the poems in which Brown found quest imagery, however. Many of these poems conjure up the image of the quest through related concepts such as voyages, lures and forms of hankering. Such concepts are not evoked systematically though a fixed set of terms. An additional complication is caused by the fact that the precise meanings of strings are often determined situationally. While the recognition of imagery functions well in most cases, the manifold semanticity of words occasionally resulted in obvious errors. The phrase "wind your gramophone", in line 22 of "An Eclogue for Christmas" was labelled as an instance of wind imagery, and the phrase "one rocks a firelit / Cradle", in lines 116 and 177 of "Flowers in the Interval" was identified as an example of rock imagery. Next to these incidental errors, which were removed manually, the software retrieved relevant references to images in the majority of cases.

6.2.11. Themes

Poems generally epitomise certain broad ideas or general emotions. Whereas the direct denotation of the words in the text can sometimes be found by making use of semantic lexicons, there is rarely a logical relationship between the tokens and the deeper themes which are being developed by these tokens. Themes can, in most cases, be found exclusively via close reading and by being sensitive to the concepts and the emotions which are evoked. In this study, no attempts were made to identify themes algorithmically. Since it was estimated, nevertheless, that it was

[439] Terence Brown, *Louis MacNeice: Sceptical Vision*, p. 108.

necessary to have data about the thematic concerns of the poems, these data were supplied manually. In *Sceptical Vision*, Brown devotes three chapters to discussions of themes. In this study, themes were correspondingly divided into three main categories: (1) romanticism, (2) a rejection of modernity and (3) a concern with metaphysics. Under these three central headings, Brown also discusses more specific themes. Using Brown's broad classification, 15 specific themes were defined. Brown's texts also contained 152 assignments of poems to these themes. The remaining poems have been classified on the basis of close readings. Assigning themes to texts is obviously a subjective task, and it is also labour-intensive. Although computation can, in some cases, accelerate research or help to make the analyses more objective, this is clearly not the case for all aspects of computer-based research.

6.3. Conclusion

In his autobiography *The Strings are False*, MacNeice explains that he frequently felt divided between two contrastive philosophical impulses. While he "wanted the world to be One, to be permanent", he also acknowledged that "any typical monistic system [i.e. based on Oneness] appeared hopelessly static".[440] This conflict between a desire to understand reality through a uniform set of principles on the one hand, and a realisation of the perplexing and diversified nature of actual phenomena on the other also forms a central concern in MacNeice's verse.[441] Poems which express a longing for permanence and for consistency, such as *Western Landscape* or *Plurality*, often contain a simultaneous acknowledgement of the notion that the world, as formulated in the well-known poem "Snow", is "incorrigibly plural". In his poem *Entirely*, MacNeice presents the world as "a mad weir of tigerish waters" (l. 19) and writes that "when we try to eavesdrop on the great / Presences it is rarely / That by a stroke of luck we can appropriate / Even a phrase entirely" (ll. 5-8). Similarly, "Train to Dublin" celebrates the diversity and the discontinuity of the phenomena that appear before the persona's mind, while simultaneously recognising an inability to recognise any structure or consistency as the poet's "half-thought thoughts divide in sifted wisps / Against the basic facts repatterned without pause" (ll. 1-2).

The pursuit of mechanisms for the computer-based detection of specific qualities of literary texts is marked by a very similar tension. Algorithms consist of a finite collocation of univocal instructions, and can consequently anticipate a limited number of cases and exceptions only. Nevertheless, they have been used in this chapter to generate metrics about literary devices whose concrete manifestations

[440] Louis MacNeice, *The Strings Are False: An Unfinished Autobiography* (Oxford: Oxford University Press 1966), p. 125.

[441] Edna Longley, *Louis MacNeice: A Study*, p. 143.

often vary deeply. MacNeice's manifold proficient experiments with repetitions of sounds or of words, for instance, have patently resulted in an expansive variety. Particular phoneme sequences are repeated entirely or in part, and they are repeated both within lines and between different lines. There are incidentally verbatim repetitions of phrases, but, as also noted by Neil Corcoran, there are also many repetitions with variations.[442] It seems virtually impossible to formulate definitive rules for the recognition of linguistic repetition in its full plurality.

Next to the challenges arising from the fact that literary phenomena may display a sheer boundless variation, this chapter has identified three additional obstacles. Firstly, it was found that there is a class of textual features whose rule-based recognition, to a large extent, remains elusive. This is the case, for instance, for the identification of allusions. It can be stipulated, for instance, that allusion is likely to occur in lines that share two or more lemmas, but it remains difficult to predict when such correspondences are of actual literary significance.[443] This assessment, in most cases, can only be made by human scholars. A second difficulty is that software often supports binary distinctions only. It assumes that a feature is either present or absent. Certain cases are clearly ambiguous, and there can be good reasons both for accepting and for rejecting a specific result. The implementation of an algorithm implies the statement of a definition of the aspect to be detected, in the very literal sense of drawing a boundary between cases which are relevant and cases which are not. It is generally difficult to create programs that can qualify statements, or that can add nuances to results. While computers necessarily reduce options to either '0' or '1', or to either black or white, research in the humanities is often intent on exploring the many shades of grey that exist between such stark opposites.

A third and obvious difficulty is that computers can only process data which are available in an explicit form, or which can be inferred unambiguously from other data which are present explicitly. As a result of this limitation, there are numerous literary devices whose unsupervised identification remains difficult. In this thesis, it was estimated, for instance, that it was impossible to extract data about themes via an algorithmic approach. Other examples include metaphor, personification, synaesthesia, chiasmus and understatement. In most of these problematic cases, recognition demands an understanding of the semantic contexts of these devices. In literary writing, the semantic context is generally too complex and too unpredictable for current semantic taggers, and, in most cases, data which require an understanding of the text's meaning can only be supplied manually.

Algorithms for the recognition of literary devices almost inevitably incur error margins. McCarty notes that imperfection is inherent to all applications in the digi-

[442] Neil Corcoran, "The Same Again? Repetition and Refrain in Louis MacNeice".

[443] Neil Coffee et al., "Modelling the Interpretation of Literary Allusion with Machine Learning Techniques".

tal humanities, as the products of the human imagination can never be restrained or tamed entirely by a single conclusive algorithm. The use of computer-based methods implies a “continual process of coming to know by manipulating things, not an achievement but an approximating convergence”.[444] Tools, more specifically, can be imperfect in three ways. Software can create a data set which is marked by a high degree of recall, but a low degree of precision. This means that the algorithms return a high number or results, many of which, unfortunately, are irrelevant. Secondly, there may be a high degree of precision, and a low degree of recall. This implies that, while the results that are returned are mostly correct, there also a large number of relevant fragments in the corpus which were not identified correctly. In the least desirable and third scenario, the software scores low both on precision and on recall. If it is indeed inevitable that tools for the identification of literary devices are structurally capricious and unreliable, a hybrid solution remains necessary, in which the computer initially produces data which subsequently need to be verified and, potentially, corrected by human scholars. This obviously places a certain limit to the scale of data sets. In the case of a low precision, scholars are confronted with a data set that contains a high degree of noise, and efforts will need to be taken to clean the results. In the case of a low recall, however, scholars will need to revisit the original sources and attempt to add data about the cases which the software had overlooked. As the removal of unwanted elements from a data set seems less labour-intensive than revisiting an entire corpus, casting the net widely and optimising recall generally seems preferable.

Terms such as ‘perfection’ and ‘imperfection’ are obviously subjective. The question whether or not an algorithm functions in a satisfactory manner can only be answered by relating its results to the expectations of individual researchers. Tools are invariably based on a prior assessment of likely forms of usage. The heuristic methods that are implemented are likewise based on assumptions and decisions that can mostly be contested. Even if an application is flawless according to one scholar’s criteria, the tool may still be inadequate for scholars with different needs. All descriptions of textual aspects require at least some level of interpretation. The algorithms that have been proposed in this chapter have evolved in sequences of trial and error. They have been calibrated and tweaked on the basis a specific collection of texts, and it is highly probable that new imperfections and new inconsistencies are exposed when they are applied to other text collections.

Despite the almost inescapable shortcomings, algorithmic processing clearly produce a number of benefits. Without digital instruments, the recognition of distinct devices such as alliteration, consonance and internal rhyme would depend fully on the alertness of individual scholars. Verse generally contains a wide range of literary devices, and it is often difficult for scholars to be alert to all possible

[444] Willard McCarty, *Humanities Computing*, p. 28.

devices simultaneously. If the detection of devices can be caught in an algorithm, scholars can produce a more encompassing description of the devices that occur in the poetry.

Digital methods are often applied to allow for explorations of corpora in their entirety. The possibility to study the whole, nevertheless, depends critically on the availability of consistent and reliable data about individual parts. Studies which are largely based on word frequencies can mostly progress directly to the level of corpora, since algorithms for the segmentation of words have been tested extensively. If data are to be created by algorithms which are still experimental, however, analyses of data sets on a macro-level are generally preceded by an evaluation of individual data values on a micro-level. The software under development generates lists of results which are potentially relevant, and the data sets that result from potentially defective algorithms mostly require further editing. This process of editing, and the evaluations of the text fragments which were selected by the software, often lead in themselves to a better understanding of the texts that are studied. Enumerations of occurrences of literary devices are comparable to a concordance. Lists of devices, taken out of their original context, can enable scholars to investigate the ways in which these terms were used throughout a body of literary works.

Scholars who are involved in the development of software for the analysis of texts can generally enhance their understanding both of the texts that are studied and of the methodology used for studying these texts. The application of algorithms, and the subsequent revision of insights, based on the results produced by these algorithms, may be regarded as a form of dialogue between reader and text, with the digital tool as an intermediary. As such, digital scholarship effectuates a redress of what Plato viewed as a crucial deficiency of written text. One of his main reasons for denouncing the written word was that it halted any discussions, as it precluded the possibility of interaction. Once a text was solidified on a static surface, the argument could no longer develop.[445] Arguably, an algorithmic engagement with the text rekindles the text's capacity to respond. The digital tool may be seen as a rendition of the scholar's understanding or conceptualisation of the source, and, provided that no coding errors have been made, running an algorithm is in effect a scrutiny of the theoretical assumptions made during the algorithmic design. The data which are produced as output may prompt programmers to reconsider these assumptions and to reassess the algorithm as implemented. Through such cycles of iterative development, scholars can theorise through practical work.

Human scholars and computers both have strengths and limitations, and these manifest themselves in opposite ways. Human scholars can be fully attentive to individual cases and to complexities discernible in particular cases. Such a focus on

[445] Plato, *Phaedrus* (Cambridge: Harvard University Press 1953).

details, however, generally hinders a comprehensive assessment of large text corpora. The approach that is taken by the machine is, in many respects, the mirror image of the method followed by the human scholar. Algorithms can collect data about the corpus in its entirety, but their observations can often be inaccurate. A systematic exploration of a text corpus mostly requires a close cooperation between the machine and the human researcher, and close reading and algorithmic processing are best viewed as complementary methods. Scholars critically face the challenge to apply computation in a dexterous manner, and to strike an astute and heedful balance between close reading and machine reading, between accuracy and completeness, and between monism and plurality.

Chapter 7

Data Representation

7.1. Introduction

In *The Art of Literary Research*, Richard Altick advises students of literature to make systematic notes of all the primary and secondary sources which are examined. Such a collection of annotations can form “a repository of factual data and ideas drawn from your sources”.[446] Using the captured notes as a basis, scholars can reflect on the various linkages and on the potential contradictions within the various facts and ideas, and such processes can eventually culminate in the formation of scholarly claims. In a study into the scholarly information practices of researchers in the sciences, the social sciences and the humanities, Palmer et al. concluded that “notetaking” can be viewed as one of the core scholarly primitives. The authors indicate that making notes can form an important part of searching and reading, and note additionally that it is often viewed as a preliminary stage in generating new original texts. [447] These findings were corroborated by Chu, who has proposed a schematic description of the scholarly process followed by literary scholars, on the basis of data acquired from structured interviews and surveys held among more than 150 researchers from different schools of literary criticism. Chu observes that literary scholars generally begin their research by producing notes about the primary and secondary materials they have selected. Such annotations point to the “quotations, answers, images, themes”[448] that are relevant in the light of a specific research question, and mostly consist of brief additional texts which clarify the relevance or the meaning of that text fragment. After the first reading, individual annotations can be related to each other, and through such associations, certain patterns may emerge. The overall argumentation generally begins to take shape when scholars become aware of certain forms of repetitions, or when they observe specific anomalies with respect to the occurrence of particular phenomena.

In a conventional setting, annotations are habitually intended for personal use only.[449] Such analogue research annotations are typically unstructured and the

[446] Richard Altick, *The Art of Literary Research* (New York: Norton 1963), p. 177.

[447] Palmer Carole L. Palmer, Lauren C. Teffeau & Carrie M. Pirmann, *Scholarly Information Practices in the Online Environment: Themes from the Literature and Implications for Library Service Development.*, p. 30.

[448] Clara M. Chu, “Literary Critics at Work and Their Information Needs: A Research-Phases Model”, in: *Library & Information Science Research*, 21:2 (January 1999), p. 260.

[449] About his bibliographic slips, Altick explains that “nobody sees those slips but me”. See Richard Altick, *The Art of Literary Research*, p. 173.

terminology is rarely fully standardised. When annotations are created digitally, however, this can often yield a number of benefits. One of the advantages is that the various facts and ideas can be searched. This advantage applies even more strongly when the annotations are captured in a structured format. When research is carried out with the aid of digital tools, this generally leads to a greater standardisation of research results. Such standardisation is typically a function of the fact that digital tools demand consistent and predictable data as input. When structured annotations can be processed algorithmically, connections, correlations and differences can often be identified more methodically.

An additional advantage of structured digital data is that these can be shared beyond the research project in which they were originally created. It is often assumed that, when researchers share some of their resources, peers can combine the various data sets that are available in order to carry out broader and more resourceful studies. In particular areas of research, such extensive data sets are indispensable. To make valid claims about developments in global literary history, for instance, it is necessary to collect details about different regions and different eras. As the vast data sets which are needed for such expansive studies can impossibly be produced by individual scholars, Franco Moretti has claimed that "quantitative work is truly cooperation".[450] When data are consolidated using a fixed format, this also enables peers to replicate the analyses that have been performed and to verify the main claims that were made on the basis of these analyses.

Open access to research data, combined with transparency about the way in which data were produced, can help to terminate the isolationist nature of humanities research and may genuinely transform it into a collaborative endeavour. Academic work in the humanities traditionally concentrates on the formation of ideas produced by individual scholars, and it is highly exceptional for a single theory or idea to be spawned by a team of researchers. Borgman notes that the humanities largely have "the lowest rate of co-authorship and collaboration".[451] Conventional scholarship frequently confirms the stereotype of the "solitary humanist; the ideal, derived from the Romantic Era, of the great mind communing with itself".[452] Computer-based scholarship, conversely, is often collaborative. [453] Tools and data sets are often constructed by interdisciplinary groups of scholars, and conferences on the use of technologies within the humanities attract increasingly large audiences. Scheinfeldt observes that scholars in the digital hu-

[450] Franco Moretti, *Graphs, Maps, Trees: Abstract Models for a Literary History* (Verso 2005), p. 5.

[451] Christine Borgman, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*, pp. 219–220.

[452] Cathy N. Davidson, "What If Scholars in the Humanities Worked Together, in a Lab?", *The Chronicle of Higher Education*, 28 May 1999, n.pag.

[453] Cathy N. Davidson, "Humanities 2.0: Promise, Perils, Predictions", in: Matthew Gold (ed.), *Debates in the Digital Humanities*, Minneapolis: University of Minnesota Press 2012.

manities are often viewed as “nice”, and that the field can be characterised using bywords such as “‘collegiality’, ‘openness’, and ‘collaboration’”.[454] The numerous online platforms on which scholars exchange ideas clearly illustrate the notion that there is a strong sense of a community.

Sharing research data also poses a number of obstacles, however. Moretti assumes that data constitute a transferrable commodity, and that, once produced, they can have an existence independently from their creator. A prerequisite for a shared use of electronic resources, however, is that these are consistently available in a standardised format, and that there can be a degree of technical and semantic interoperability. Standardisation does not always seem possible in the case of humanities research. Studies usually focus on the myriad of ways in which human beings have expressed themselves, and the interpretation of such highly diverse artistic utterances cannot always be formalised unequivocally. In addition, standardisation inevitably imposes certain limitations. When standards are being developed, the aim is usually to describe a domain and to allow room for as many aspects as possible. Once a standard is adopted, it loses some of it flexibility and users of the standard may need to adjust their own descriptive practices to the standard. A standard also restricts what can be said and how things can be said, and full expressiveness may need to be sacrificed for the higher goal of being able to collaborate.

Despite these potential difficulties, structured annotations can be valuable for a variety of reasons, and, within the case study that was conducted for this thesis, a decision was also taken to capture all the generated data using a standardised data format. As part of the case study, a software application was developed which can generate annotations about literary phenomena such as rhyme, metre, assonance and alliteration. This application has been used to produce secondary data about the poetry of Louis MacNeice. In total, 246.660 observations have been generated about the 311 poems that were analysed. As it seemed crucial to ensure that these observations could all be captured accurately, the selection of the data format was based on three requirements. Firstly, the format needed to provide support for a sufficiently rich ontology which can accommodate the various phenomena this study has focused on.[455] A second requirement was that it needed to be possible to connect the terms from the ontology to specific text fragments. The representation language in itself could not impose any practical limitations in connecting terms from the ontology to specific text fragments. Thirdly, when a term was connected to a text fragment, it also had to be possible for a scholar to claim responsibility for this act. The act of describing a literary text is, almost inevitably, interpretative. Systems for the representation of data ought to be based on the assumption that

[454] Tom Scheinfeldt, “Why Digital Humanities is “Nice””, in: Matthew K. Gold (ed.), *Debates in the Digital Humanities*, Minneapolis: University of Minnesota Press 2012, p. 59.

[455] For a discussion of the term ontology, see section 4.2 of this thesis.

there are no objective facts about human artefacts in themselves. The only class of facts that may be said to exist is that a certain scholar, working at a particular moment in time, made a statement about a particular artefact. Data formats must be capable of recording such controversy.

As was also discussed in Chapter 6, the data that were produced within the case study have partly been captured using The Text Encoding Initiative (TEI). The standard provides a valuable method for describing the logical structure of the text, amongst other aspects. Because it was found, however, that the standard did not fully meet the three requirements which were formulated above, the majority of the observations about MacNeice's poems have been recorded using the Open Annotation Collaboration (OAC). This chapter offers a motivation for the two data formats which were chosen and additionally discusses the strengths and the shortcomings of these two formats. Importantly, the description of the weaknesses of TEI in section 7.2 differs in a qualitative sense from the discussion of the affordances of OAC in section 7.3. TEI has already been in use for more than 25 years,[456] and the practical details of its use have already been discussed extensively elsewhere. The criticism of the standard is mostly of a fundamental and theoretic nature. OAC, on the contrary, is currently an emerging technology, and there are still very few texts on the ways in which the protocol can be used, especially in the domain of literary studies. The discussion of the strengths and weaknesses of OAC is, for this reason, much more practical and concentrates more specifically on the techniques that have been applied in the case study of this thesis.

7.2. The Text Encoding Initiative

Unsworth has explained that digital models are based on ontologies.[457] The suitability of digital surrogates depends for a large part on the nature of the underlying ontology. To gauge the utility of the TEI for the case study in this thesis, it is necessary to consider the standard's facilities for the description of textual aspects that are typically associated with poetic texts.[458] The <l> and <lg> elements, both from the TEI core set, may be used to mark up the structural division of a poem into stanzas and into verse lines. Within <lg>, the @type attribute can be used, so that a distinction can be made between stanzas of different lengths. Both <lg> and <l> may be combined with a number of attributes that can offer information on rhyme schemes and metrical patterns. More specifically, the @met attribute can be

[456] James Cummings explains that "The TEI grew out of a recognized need for the creation of international standards for textual markup that resulted in a conference at Vassar College, Poughkeepsie, in November 1987." See James Cummings, "The Text Encoding Initiative and the Study of Literature", p. 451.

[457] John Unsworth, "What Is Humanities Computing and What Is Not?".

[458] Most of these are described in section 6 of the P5 TEI guidelines, <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/VE.html> (16 March 2013)

used to record either a term descriptive of the metrical pattern (e.g. "iambic pentameter") or a schematic representation of the metre (e.g. "-X-X-X-X-X"). The @rhyme attribute is generally used at the level of the <lg> element to indicate the rhyme scheme of the lines contained within the line group. On the individual lines, the words that rhyme may potentially be encoding using the <rhyme> element. On the basis of this mechanism, both final rhyme and internal rhyme may be recorded.[459] Importantly, since the @rhyme attribute cannot be used more than once within an <lg> element, encoders can only provide information about a single form of rhyme. Many of MacNeice's poems contain adroit combinations of perfect rhyme, slant rhyme or semi-rhyme, however. Examples include "The Habits", "Homage to Clichés" and "Solstices". By default, the TEI standard offers insufficient support for the description of such concurrent instances of different forms of rhyme.

While some aspects of metre and of rhyme may be captured within attributes of the <l> and the <lg> elements, the TEI crucially lacks specialised vocabulary for the description of figures of speech or of figures of thought. About the annotation of figurative language, the guidelines propose that, given "the great richness of modern metaphor theory", any single proposal "would have seemed objectionable to some and excessively restrictive to many". The guidelines suggest that, when particular descriptive terms are not supplied by the standard, encoders ought to add these terms themselves. During the development of the Myopia application, which is a visualisation environment aimed to support the close reading of poetry, Chaturvedi et al. have addressed these limitations of the TEI via the definition of bespoke elements and attributes for particular literary phenomena.[460] The elements <ambiguity> and <connotation> were created, firstly, to enable scholars to record particular words associated with the tokens in the poetry. Figurative language could also be described via a <metaphor> element with associated @tenor and @vehicle attributes. To identify verse feet, a <foot> element was added, and the syllables contained in these could be categorised as iambic, spondaic and pyrrhic patterns via the @type attribute. <subject>, <verb> and <object> were added to record the syntactic structure of sentences. Figures of speech based on acoustic patterns could be characterised using <consonance>, <assonance>, and <alliteration>.

An alternative approach, next to the creation of additional terms in customised TEI schemas, is to make use of the existing <interp> element. This element may be used to associate a user-defined descriptive term with a unique identifier. This identifier can subsequently be used as the value of the @ana attribute. Amongst many other elements, this attribute may be used within a <seg> element, which

[459] The TEI P5 guidelines coincidentally use Louis MacNeice's poem "The Sunlight on the Garden" to illustrate the facilities for encoding occurrences of internal rhyme.

[460] Manish Chaturvedi et al., "Myopia: A Visualization Tool in Support of Close Reading", (2012), n.pag.

can be used to delineate a specific sequence of words, without immediately specifying the nature of the selected text fragment. Kate Singer has argued that such expansiveness is very beneficent from the viewpoint of literary criticism. Within the TEI standard, there is clearly a dearth of terms for the description of literary devices. Such paucity, however, stimulates encoders to "interrogate poetics terms and interpret texts in descriptive rather than prescriptive ways precisely because they do not automatically resort to classical literary terminology".[461] Descriptive encoding also implies a form of classification. Singer speculates that, if the TEI consortium had supplied a highly developed system for the classification of literary devices, encoders would have been prompted to apply these terms mechanically, and to overlook phenomena which lie outside predefined categories. She surmises additionally that the ability to supply new terms via the @ana attribute encourages scholars to reflect critically on the unique properties of specific fragments, and to devise singular terms to describe their qualities. Such new terms may describe the connotations of specific words or they may characterise the nature of recurrent tropes. The terms that are used traditionally with literary criticism are, to a large degree, historically contingent, but the TEI also enables scholars to supply local or idiosyncratic variants for the received terminology and to effectuate a degree of acclimatisation.

Whereas TEI encoding can in theory be added both manually and via text mining algorithms, the mark up is incorporated most frequently via the former method. The encoding is commonly the reflection of a close reading of the text. Close reading usually consists of deeply attentive forms of engaging with texts, and readers may respond to the words they encounter in widely diverse ways. It may be argued that the technical structure of the TEI mirrors such variability. Via its lenient ontology and via its extensibility, the TEI offers support for idiosyncratic and irregular ways of interacting with texts. In quantitative analyses of large collections of texts, however, it is generally important to explore the broader patterns and to establish the interconnections between distinct texts. For such applications, it is essential to ensure that related phenomena have been identified identically throughout the entire corpus. In many cases, such consistency can only be achieved if terms have been described via an ontology which has been applied strictly.

The TEI Schema, which is maintained by the Text Encoding Initiative Consortium, may be considered a "loose" ontology, in two important senses. Firstly, the standard clearly does provide a vocabulary which can be used to describe and to encode specific characteristics of texts. In the introductory chapters of the TEI guidelines, it is stated that the standard is based on an "abstract model" and that each element has a specific "semantic function" which encoders ought to

[461] Kate Singer, "Digital Close Reading: TEI for Teaching Poetic Vocabularies", in: *Journal of Interactive Technology and Pedagogy*, 3, n.pag.

respect.[462] In a formal and strict ontology, relations between entities have been defined explicitly.[463] A limitation of the TEI standard, and of XML-based languages for the addition of in-line mark up in general, is that the semantic relations that can exist between elements cannot be stated explicitly. DTDs and XML Schemas can specify that one element ought to be contained or nested within another element, but no information can be provided on the precise meaning of this nesting. Renear et al. observe that all XML-based markup languages share a degree of ambiguity. There are no facilities for the description of "the fundamental semantic relationships amongst document components and features in a systematic machine-processable way".[464] The intended meaning of a specific element cannot be inferred from the way it is used in the document, and knowledge about the intention of these elements must be built into the software applications that query these texts. Willet notes that "SGML/XML markup itself is not a "data model", as it exclusively "serializes a data structure".[465] The technique may be viewed as controlled vocabulary, rather than as a formal ontology.

Secondly, the TEI ontology may also be viewed flexible or loose because of its extensibility and its leniency. Encoders are empowered to modify existing elements and attributes and to add new terms according to their own needs, thereby undermining the concept of a central ontology. In an important sense, the TEI enables scholars to give expression to a phenomenology, as elements frequently reflect both a scholar's idiosyncratic understanding of a descriptive category, and the subjective views on the textual aspect to which the category is applied. Research projects which have identified lacunae in the guidelines or in the technical possibilities of the TEI standard are often forced to develop project-specific elements and attributes. If the needs for extensions to the standard are sufficiently widespread, SIGs can be established to institutionalise ontological views of particular sets of phenomena. In spite of the fact that the TEI endorses specific ontological commitments, there is generally no widespread consensus on how and under which circumstances specific elements should be applied. Since the aim from the onset had been to accommodate hugely diverse research areas, the standard provides "a maximum of comprehensibility, flexibility, and extensibility".[466]

[462] "TEI Guidelines for Electronic Text Encoding and Interchange", <http://www.tei-c.org/Guidelines/P5/> (12 March 2013)

[463] Lee Lacy, *OWL: Representing Information Using the Web Ontology Language*, p. 36.

[464] Allen Renear, David Dubin & C. M. Sperberg-McQueen, "Towards a semantics for XML markup", in: *Proceedings of the 2002 ACM symposium on Document engineering - DocEng '02*, (New York, New York, USA: ACM Press, 2002), p. 119.

[465] Allen H. Renear, "Text Encoding", in: *A Blackwell Companion to Digital Humanities*, Oxford: Blackwell 2002, p. 237.

[466] TEI Guidelines for Electronic Text Encoding and Interchange, <http://www.tei-c.org/Guidelines/P5/> (12 March 2013)

The idiosyncrasy of encoding practices, and the extensibility of the vocabulary can obviously jeopardise the reuse of encoded texts beyond the project in which they were originally created. This aspect is not necessarily problematic when cross-project interoperability is not an objective, and when scholars principally aim to capture the annotations about texts in a systematic manner for the purpose of their own research. In some cases, however, the data format itself may also hinder an effective representation of research annotations. One particular problem inherent in XML-based encoding systems is that the format may lead to conflicting hierarchies. The W3C recommendations stipulate that XML documents must follow a strict hierarchical structure. Documents, more specifically, ought to have exactly one root element, and all other elements need to be contained hierarchically within this single root.[467] Various theorists of textuality correspondingly surmised that texts are typically composed of distinct constituent parts, which are related to each other in a hierarchical manner. This theory views texts as "Ordered Hierarchies of Content Objects" (OHCO).[468] A poem, for instance, may be composed of stanzas, and these stanzas, in turn, may consist of lines. Various authors have also argued that the OHCO theory is flawed, since there are also numerous examples of textual phenomena which do not nest neatly. Collisions of hierarchies occur, for instance, in verse texts which contain enjambments. The verse lines may be encoded using the <l> element, but, when scholars also want to encode grammatical sentences using <seg>, this evidently results in invalid XML. Phenomena that appear in poetry, such as the tropology, the syntax, the metre and the literary devices based on sound frequently overlap, too. A metaphor, for instance, may start at one line, and end on the line that follows.[469]

Renear et al. have argued that the OHCO theory may be redeemed if it is accepted that a text may be seen as "system of concurrent perspectives which decompose into concurrent sub-perspectives which in turn can be decomposed".[470] Physical, prosodic or syntactic analyses of a text all produce distinct hierarchies, but, since content objects within a single hierarchy always seem to nest properly, an overlap of content objects should be taken as an indication of the fact that these two objects belong to different analytic perspectives. For the purpose of the Myopia viewer, for instance, the scholars produced four separate files, and each of these focused on a separate hierarchy. It can easily be envisaged that duplicating a source text multiple times introduces difficulties when this source text, for whichever

[467] W3C, "Extensible Markup Language (XML) 1.0 (Fifth Edition)", 2008, <http://www.w3.org/TR/2008/REC-xml-20081126/> (16 July 2014)

[468] Allen Renear, David Durand & Elli Mylonas, "Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies", in: *Research in Humanities Computing*, Oxford: Oxford University Press 1995.

[469] Manish Chaturvedi et al., "Myopia: A Visualization Tool in Support of Close Reading".

[470] Allen Renear, David Durand & Elli Mylonas, "Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies", n.pag.

reason, needs to be updated. An alternative approach is to select a single hierarchy as the primary hierarchy within a file, and to flatten other hierarchies, by using milestone tags, or elements without a body. This approach, however, can only be effective when the data to be recorded consists solely of the identification of a particular location within the text. This may be the case for a page break or a line break. When specific words actually need to be delineated and categorised, as in the case of assonance or of metaphors, the problem of conflicting hierarchies persists.

In his article "Digital Representation and the Text Model", Dino Buzzetti draws attention to an additional obstacle inherent in descriptive mark up. Buzzetti distinguishes between the expression and the context of the text. The former term is "the linear order of the succession of codified characters" and the latter term is used to refer to "that which the various strings of characters signify".[471] Cesar Segre clarifies that literary theorists have variously used the term 'content' to refer to "themes, composition, and genres" or to "ideas, feelings and inspirations".[472] These concepts or sentiments collectively form a "semiotic product"[473] which critics may extract from the concrete expression by evaluating, for instance, the connotative effects of words. The structure of the content, however, does not necessarily coincide with the structure of the expression. In many cases, the central ideas of a text cannot be connected unequivocally to particular sequences of characters. Data on the meaning of an extended metaphor, the general setting of a text, the central thematic concerns, or the intertextual connections to other texts cannot always be captured using a standard which focuses exclusively on the linear string of characters. MacNeice's poem "Charon", for instance, is basically an extended metaphor in which a bus trip across London is portrayed as a voyage across the river Styx. The hands of the bus driver are "black with money", his "eyes are dead" and the passengers also note his "varicose veins". Various words in the poem collaborate to produce a nightmarish effect. Brown argues that the repetition of the phrase "we just jogged on" conveys "the irreversible process of life towards death".[474] This central subtext of a death in life, however, cannot be associated exclusively to a single sequence of words within the text. Since the extended metaphor cannot be connected to a single vehicle, it cannot be encoded effectively using TEI-based mark up.

Next to the difficulties that result from multiple hierarchies, and from the fact that occurrences of phenomena cannot always be situated in a single location within the text, a third difficulty that inheres in TEI-based encoding is that descrip-

[471] Dino Buzzetti, "Digital Representation and the Text Model", in: *New Literary History*, 33:1 (2002), p. 68.

[472] Cesare Segre, *Introduction to the Analysis of the Literary Text* (Bloomington: Indiana University Press 1988), p. 41.

[473] Dino Buzzetti, "Digital Representation and the Text Model", p. 88.

[474] Terence Brown, *Louis MacNeice: Sceptical Vision*, p. 68.

tive terms can generally be applied only once. As was discussed, the @ana attribute may be used to record, for instance, a term indicating the connotation of a word. Once a word or a sequence of words has been marked up, it cannot be encoded at the same hierarchical level with another element. Within a single file, any decision to apply a particular element effectively suppresses all opposing views. Many of the statements that are made via the TEI are of a subjective nature, however. Huitfeldt argues that "[t]here are no facts about a text which are objective in the sense of not being interpretational". A scholar encodes the text "in accordance with his or her interpretation", and "the transcriber's interpretation is not theory-independent". [475] Eggert writes likewise that "every electronic representation of text is an interpretation". Encoding is "doomed to remain problematic, incomplete and perspectival".[476] Aspects pertaining to conflicting hierarchies cannot be recorded within a single TEI file, and the same is true for the registration of disagreement. In XML documents, it is generally arduous to express multiple equivalent opinions about a single fragment. A particular encoding monopolises a single opinion, and scholars who wish to record alternative perspectives need to create a new manifestation of the text. In Chapter 3, plain texts and annotations about texts were described as different types of data, but, in a TEI file the two typed of data are intermingled in a single file. Within this format, a new set of annotations can only be created by duplicating the full primary data, implying an inefficient use of primary data.

In Chapter 3, a distinction was also discussed between annotations which describe aspects which are explicit and observable, and annotations which describe implicit textual aspects. Data in the latter category are often speculative and interpretative. Meister stresses that "interpretation is an interpretation if and only if at least one alternative to it exists".[477] When tags are inserted that characterise the use of metaphors, or that locate instances of other literary devices such as onomatopoeia or synaesthesia, the files generally reflect the interests and the interpretations of one particular scholar. Describing the semantic contents of the text is crucially an open-ended process, as scholars may interpret and re-interpret their sources virtually limitlessly. For this reason, it seems inadequate to incorporate such observations directly within the primary text. Annotations about the logical structure or the typography, conversely, possess a degree of objectivity, as they can be verified via a consultation of the original sources.[478] As observations

[475] Claus Huitfeldt, "Multi-Dimensional Texts in a One-Dimensional Medium", in: *Computers and the Humanities*, 28 (1994), pp. 237–239.

[476] Paul Eggert, "The Book, the E-Text and the "Work-Site"", in: Marilyn Deegan & Kathryn Sutherland (eds.), *Text Editing, Print and the Digital World*, Farnham: Ashgate 2009, pp. 67–73.

[477] Jan Christoph Meister et al., "Crowdsourcing meaning: a hands-on introduction to CLÉA, the Collaborative Literature Éxploration and Annotation Environment", in: *Digital Humanities 2012*, (Hamburg: 2012), n.pag.

[478] The observable typographical features of a text may be described though encoding, but their meaning may still be open to interpretation.

about the logical structure of the text may be assumed to be relatively factual and stable, proximate inclusion into the plain text seems less problematic.

The TEI can reasonably be used for descriptive annotations about the logical components of the expression, but the standard seems less suitable for the consolidation of conjectures about the text's content, as these are often contentious and idiosyncratic. Subjective and interpretative observations can arguably be captured more appropriately in a separate document and independently from the document that contains the full text. To some extent, such a division was implemented in the Just in Time Markup (JITM) project, which was developed at the Australian Scholarly Editions Centre at the Australian Defence Force Academy. In this project, the type of separation that was effectuated however, was more rigorous. Berrie explains that, in the JITM system, the bare transcriptions are stored separately from all scholarly annotations of these texts. At the request of a user, a specific set of mark up tags can be added to a transcription, thus allowing the creation of "on demand user-customised versions of electronic editions".[479] Since the interpretative mark up is not inserted into the actual transcription file, this latter resource remains authentic. It also ensures that transcriptions can be reused. Marked-up files generally privilege a particular perspective on the text, but if the mark-up is recorded separately from the text, it also becomes possible to record opposing views on the text. The transcription file, once completed, is assumed to remain static, and a reference scheme is used to insert mark up into the text. As a result, "the electronic edition can become an evolving work of scholarship based on the work of many hands".[480] In JITM, transcription files are tokenised by making use of the spaces that appear in the document. The spaces are assumed to delineate words, and these words in turn consist of characters. The file which records the mark up works with position codes such as "17.001" or "28.006". In this particular case, these codes refer to the first character of the 17th word or the sixth character of the 28th word. Mark up codes can be inserted dynamically at the positions which are recorded in this manner. This system of capturing positions appears to be fragile, as positions may evidently change when the transcription file changes at some point.

A related form of stand-off mark up was implemented in the CATMA application, which was developed at the University of Hamburg. CATMA is described as "a practical and intuitive tool for literary scholars, students and other parties with an interest in text analysis and literary research".[481] Scholars can initially supply a collection of TEI-encoded texts. Within the tool's interface, scholars can select specific fragments and describe these using their own tags. These tags are stored in a separate *User Mark up* document. This document contains references to the

[479] Phillip William Berrie, *Just In Time Markup for Electronic Editions*, n.pag.

[480] Ibid.

[481] <http://www.catma.de/> (12 April 2014)

fragments these tags are applied to. As the descriptive tags and the tags that explain the logical structure of the text are separated, there is no risk of overlapping or contradictory mark up. Several User Mark up files can be created for a single source document. Tags can also be structured hierarchically, as users can create tags and "subtags". When a user searches for a "supertag", the subtags are retrieved as well. One difficulty is that the tags are tied to the CATMA application, and that these cannot be used outside of the environment. It is also difficult to ensure cross-project interoperability. The functionalities of the CATMA tool are further expanded within *Collaborative Literature Éxploration and Annotation*,[482] which aims to "supplement Google Books with a web based collaborative text exploration, markup and analysis environment".[483]

7.3. Semantic web

The Worldwide web has been described as "the most powerful communication medium the world has ever known".[484] It is used by millions of people on a daily basis, it determines much of our social and our cultural lives, and it is also one of the largest drivers of our global economy. Tim Berners-Lee, who is commonly viewed as the progenitor of the Web, has explained that the current ubiquity of the Web was largely the result of its flexibility and of its decentralised nature. Berners-Lee recognised that the web would never have developed into the popular global phenomenon that it is today if it had imposed strict and rigid rules on how to structure web pages and on how to create hyperlinks. To ensure a widespread adoption, "the Web had to throw away the ideal of total consistency", and, on the Web, "anything can link to anything".[485] The Web generally lacks a formal and consistent structure, and this means that its contents cannot be searched in the same way a well-designed database can. The effectiveness of searches is often hampered by the fact that the content consists of texts which were designed for human readers. Web pages are typically created using HTML, which is a standard mainly for the specification of the typographic appearance of web pages. This focus on presentational aspects jeopardises the accessibility of the data that is contained in these documents, as search engines can usually carry out full text searches only.

[482] <http://www.catma.de/webfm_send/22> (12 April 2014)

[483] Jan Christoph Meister et al., "Crowdsourcing Meaning: A Hands-on Introduction to CLÉA, the Collaborative Literature Éxploration and Annotation Environment".

[484] World Wide Web Foundation, "History of the Web", <http://webfoundation.org/about/vision/history-of-the-web/> (14 May 2013)

[485] Tim Berners-Lee, James Hendler & Ora Lasilla, "The Semantic Web: A New Form of Web Content That Is Meaningful to Computers Will Unleash a Revolution of New Possibilities", in: *Scientific American*, (2001), pp. 73–74.

Since the current web predominantly consists of "multimedia human-readable material", it basically functions as a "glorified television channel".[486]

The Semantic Web consists of a collection of techniques which aim to ensure that the content on the Web can be processed more systematically by machines. Berners-Lee (2002) explains that the Semantic Web was already part of the original vision for the web as it was conceived in the late 1980s. The techniques that were envisioned aimed to "bring structure to the meaningful content of Web pages".[487] The objective of the semantic web was not to replace the current web, but to extend it with an additional layer through which information can be given a "well-defined meaning, better enabling computers and people to work in co-operation". The aim of the Semantic Web is to publish those structured data collections directly on the Web, thus precluding the need to translate these data into human-readable HTML files first. Using semantic web techniques, machines can exploit data which is currently "in relational databases, XML documents, spreadsheets, and proprietary format data files".[488] The Semantic Web makes use of a number of central components. Concepts, importantly, need to be referred to using identifiers, rather than via words. The Semantic Web makes use of the general web architecture, which, according to W3C, is "an information space in which the items of interest, referred to as resources, are identified by global identifiers called Uniform Resource Identifiers (URI)". A second central component of the Semantic Web is the *Resource Description Framework* (RDF).[489] URIs and RDF can be used in combination to describe concrete and abstract concepts, together with the manifold relationships that can exist between these concepts.

RDF offers a framework which can be used to make claims about a particular domain, and, as such, it can serve as an alternative to the TEI. To create structured data about text fragments, it is necessary, as a first step, to unambiguously delineate and identify the fragments which need to be annotated. To be able to include these fragments into an RDF assertion, they additionally need to be defined as URIs. Joel Kalvesmaki explains that, throughout the history of textual scholarship, systems of canonical references have been used to label specific sections of works. These references enable scholars to cite particular parts of a text, without having to indicate a particular edition.[490] Large works such as *The Bible* or Homer's *Illiad*, for instance, have been segmented into units which can be referred to separately using codes such as *1 John 4:19* or *Homer, Iliad 1.1*. This system of

[486] Tim Berners-Lee, "Foreword", in: Dieter Fensel (ed.), *Spinning the Semantic Web*, Cambridge: MIT Press 2005, pp. xi–xiv.

[487] Tim Berners-Lee, James Hendler & Ora Lasilla, "The Semantic Web: A New Form of Web Content that is Meaningful to Computers will Unleash a Revolution of New Possibilities", n.pag.

[488] Ibid.

[489] RDF has been discussed earlier in Chapter 3.

[490] Joel Kalvesmaki, "Canonical References in Electronic Texts: Rationale and Best Practices", in: *Digital Humanities Quarterly*, 008:2 (2014), 1.

canonical references can usefully be extended into the digital realm. Kalvesmaki stresses, however, that there are currently no widely accepted interoperable protocols for defining and resolving canonical references. In texts that are encoded using the TEI, individual units can be associated, for instance, with a canonical reference using the @cRef attribute. These references can subsequently be incorporated in a URI. To ensure that references can function correctly, the server which hosts the TEI file that contains these fragments needs to be able to return the associated fragment. An alternative is to make use of the *Canonical Text Services* (CTS) protocol. The protocol defines a URN scheme in which the work and the fragment can both be identified. The identifier for the passage is generally derived from an identifier that is specified in a TEI document. To ensure the uniqueness of the URNS, a CTS URN also demands a declaration of a CTS URN namespace.

Since, within the current study, no authoritative CTS URN namespace had been registered yet, a decision was taken to base the referencing system on the URLs of the online files. In the case study, all poems have been encoded in TEI. The TEI encoding focuses mostly on the logical structure of the texts, and the different stanzas, lines and words have all been given unique identifiers. A URI may be created by combining the URL of the poem with a unique fragment identifier, derived from the values of the @n attributes within the TEI files. The second line of the poem "Belfast", for instance, may be referenced using the URI `<http://www.bookandbyte.org/macneice/belfast.xml#l2>`. The same principle can be used to refer to individual words, as these have been numbered similarly.

Ranges of words may be referenced by making use of the XPointer standard. XPointer offers a provision for the identification of longer fragments of texts, in which the identifiers of the opening point and the closing point need to be supplied.[491] This technique can be applied usefully for the description of instances of enjambment, for instance. One example of this device can be found in the second canto of *Autumn Journal*: "But tonight is quintessential dark forbidding / Anyone beside or below me". To refer to the full sentence, the following URI may be constructed: `<http://www.bookandbyte.org/macneice/AutumnJournalII.xml#xpointer(id("w123")/range-to(id("w145")))>`. In this example, "w123" is the value of the @n attribute which was assigned to the first word in the grammatical sentence. In this study, it was assumed that the individual word constitutes the smallest unit that needs to be addressed. If, in other studies, it is necessary to comment on individual characters, an alternative solution may be chosen, in which the TEI encoding also assigns identifiers to characters.

The descriptive terms that are used in a Semantic Web application are generally derived from a series of ontologies. A broad variety of ontologies have

[491] <http://docstore.mik.ua/orelly/xml/xmlnut/ch11_07.htm> (21 June 2014)

been made available already, but new terms can be created using the Web Ontology Language (OWL). OWL, more specifically, provides a mechanism for describing particular classes of objects. The technique can be used to mint identifiers for such classes, and to record some of their formal characteristics. The declaration of a class can also include a textual definition of the term, which may enable human readers to evaluate its suitability.[492] When the OWL-based ontology is made publicly accessible, the classes that have been declared can be used within RDF assertions.

Within the case study that was conducted in this thesis, data have been produced about a wide range of literary devices such as alliteration, assonance, onomatopoeia, rhyme and metre. Since it was found that the majority of the phenomena investigated have not been described yet in any existing ontology,[493] a new ontology of literary terms was developed, using OWL. The classes “Poem”, “Stanza”, “Line” were created to characterise the structural components of poetic texts. The classes “Couplet”, “Tercet”, “Quatrain”, “Quintain” and “Sestet” were established to enable scholars to distinguish between stanzas of different lengths. Importantly, classes were also defined for the description of the literary devices that were investigated in the case study.[494] Terms were defined more specifically, for “perfectRhyme, “AssonanceRhyme”, “consonanceRhyme”, “SemiRhyme”, “Alliteration”, “Assonance”, “Consonance”, “InternalRhyme”, “DeibhideRhyme”, “AicillRhyme” and “InternalConsonanceRhyme”. It was also specified that assonance rhyme and consonance rhyme are both specific cases of slant rhyme. In the human-readable descriptions of these terms, definitions were cited from the *Princeton Encyclopedia of Poetry and Poetics*, and from Abram’s *Glossary of Literary Terms*. The generic class “Rhyme” was declared, additionally, to be able to cluster the various more specific forms of rhyme. The number of terms and the types of relations between these terms may be expanded at a later stage. It would be possible, for instance, to define a hierarchical structure in which figures of speech are distinguished from figures of thought. For the purpose of the current case study, however, a declaration of the basic terms was considered sufficient.[495]

Next to having a suitable ontology, a method needs to be chosen to connect terms from this ontology to selected text fragments. Such structured annotations can be created using a number of technologies. This section will focus on the

[492] Lee Lacy, *OWL: Representing Information Using the Web Ontology Language*, p. 173.

[493] Searches for these terms in the linked open data search engine *Falcons*, available at <http://ws.nju.edu.cn/falcons/objectsearch/index.jsp >, did not produce any results.

[494] The ontology that was created followed the classification that was discussed in Chapter 2 of this thesis.

[495] Information about the full ontology that was implemented can be found in Appendix B.

nanopublications protocol and on the *Open Annotation Collaboration* (OAC).[496] Nanopublications, firstly, were originally conceived of by members of the Concept Web Alliance.[497] The technique entails a mechanism for making individual academic findings available separately, in a machine-readable format. A nanopublication is viewed as the "smallest unit of publishable information".[498] The technique may be used to publish "quantitative and qualitative data, as well as hypotheses, claims, and negative results that usually go unpublished".[499] Since each nanopublication is made available under a single URI, they can also be cited, for instance, in textual publications. The concept of nanopublications was largely developed in response to a perceived information overload. Textual scholarly publications typically contain a multitude of statements expressed in natural language, and the sheer quantities in which such textual publications are made available at present clearly complicate staying abreast. Findings which are made available as nanopublications may be processed efficiently by machines, and Mons and Velterop expect that computer applications can produce "real time alerts"[500] for the benefit of researchers interested in specific concepts.

A nanopublication, more concretely, can consist of three components. The central component is the assertion. It is an RDF triple which represents the main finding that is made available. Since it was recognised that findings frequently need to be understood within a specific context, a second component was defined in which researchers can describe the conditions under which the assertion is considered to be true or relevant. Thirdly, and importantly, the provenance of the assertion can also be captured. In this third section, data can be supplied about the creator of the assertion, i.e. the person who is responsible for its intellectual contents, and about the time and date on which the assertion was made. Data about the provenance should enable peers to evaluate the trustworthiness of the statement. One notable characteristic of the provenance section is, additionally, that researchers may record the type of evidence. Velterop and Mons discuss a number of evidence types. The assertion that is captured in a nanopublication may be "derived from observation or measurement". Alternatively, it may be "derived as

[496] Sanderson et al. note that there have been a number of other technologies for the registration of structured annotations, including Annotea, Hypothesis and iAnnotate. Many of these technologies have shortcoming or are no longer continued. The two technologies which are discussed in this section are considered to be the most appriate solutions in the academic domain. See Robert Sanderson, Robert Sanderson, Bernhard Haslhofer, Rainer Simon, et al., "The Open Annotation Collaboration (OAC) Model", <https://arxiv.org/pdf/1106.5178.pdf>

[497] The CWA is "an open collaborative community that is actively addressing the challenges associated with the production, management, interoperability and analysis of unprecedented volumes of data"

[498] Erik Schultes & Mark Thompson, "Using Nanopublications to Incentivize the Semantic Exposure of Life Science Information", p. 1.

[499] <http://nanopub.org/guidelines/working_draft/> (2 November 2014)

[500] Barend Mons & Jan Velterop, "Nano-Publication in the e-science era", (2009), n.pag.

a prediction based on a model or theory”.[501] In this current case study, recording the evidence type is highly relevant. Using this provision, a distinction can be made between data which were generated algorithmically on the one hand, and data which have been created or edited manually on the other.

The code below gives an impression of how the nanopublication framework can be applied to express the observation that the second line of Louis MacNeice’s poem “Belfast” contains alliteration.[502]

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix nanopub: <http://www.nanopub.org/nschema#>.
@prefix nphbvar: <http://www.nanopub.org/nanopubs/hbvar#>.
@prefix dcterms: <http://purl.org/dc/terms/>.

{
   nphbvar:n0 nanopub:hasAssertion nphbvar:n0assertion.
   nphbvar:n0 nanopub:hasProvenance nphbvar:n0provenance.
   nphbvar:n0 rdf:type nanopub:Nanopublication.
}

nphbvar:n0assertion {
       http://www.bookandbyte.org/macneice/#l2 lit:contains
       lit:Alliteration.
}

nphbvar:n0provenance {
       nphbvar:n0assertion
       dcterms:created ‟2014-12-12”^^xsd:date.
       nphbvar:n0assertion nanopub:authorID info:eu-
       repo/dai/nl/32959737X.
}
```

As can be seen, this example does not make use of the context component. In general, this component can be used to qualify a statement, and to explain that the claim applies exclusively under specific conditions. This provision, however, does not seem applicable to largely subjective statements about literary texts. The provenance section, however, is highly germane, as it offers a solution to one of the crucial limitations of the TEI format. If there is disagreement about the observation

[501] Barend Mons & Jan Velterop, “Nano-Publication in the e-science era”, n.pag.

[502] The RDF graph is represented via the Turtle notation.

that is made, contrastive claims can be published as separate nanopublications. Each alternative claim can be timestamped and attributed individually.

The nanopublication framework also presents a number of difficulties. As was discussed, the method that was developed within the case study for the recognition of alliteration does not only signal the occurrence of this device. It also produces a pattern representing the sounds that are repeated. Crucially, since a nanopublication can only be associated with one assertion,[503] it does not seem possible to record this pattern, in addition to the fact that the passage contains alliteration in itself. In some cases, it also seems difficult to describe literary phenomena exclusively via a tripartite structure. Lines 47 and 48 of the eleventh canto of "Autumn Journal" contain an instance of aicill rhyme: "Who know that truth is nothing in abstraction / That action makes both wish and principle come true". Since the individual words were marked up using the <w> element, and since each <w> element has been assigned an @n attribute, words can be identified separately. The subject of this assertion, nevertheless, consists of two separate parts. The words "abstraction" and "action" collectively produce the internal rhyme, but the nanopublications framework does not allow for assertions which contain multiple subjects.

A number of these difficulties can be addressed by making use of the OAC. It is a protocol that can be used to capture structured annotations.[504] The framework enables annotators to express a "relationship between two or more resources, and their metadata". The OAC proposes a framework in which system-independent annotations can be published in accordance with linked open data standards. The OAC data model propose three central classes. The "Annotation" itself consists of a "Target", which represents the object that is annotated. The "Body" is a comment or any other resource which offers information about this target. OAC is primarily a method for the management of annotations, and its data model does not propose any terms beyond those terms that are needed to construct the annotation itself, and to describe its provenance.

The descriptive terms to be used in the annotation are not stipulated in the model itself, however.[505] In the OAC guidelines, it is explained that scholars can classify particular text fragments using terms derived from an external ontology. This is viewed, more specifically, as an example of semantic tagging. In OAC, the ontology terms can be supplied directly within the body of the annotation, and this tag must be associated with the class oa:SemanticTag. The OAC guidelines also encourage annotators to explain the reasons for creating an annotation, using the property a oa:motivatedBy. Within the context of this current study, the term

[503] <http://nanopub.org/guidelines/working_draft/> (18 September 2013)

[504] <http://www.openannotation.org/spec/core/> (18 September 2013)

[505] Dirk Roorda & Charles van den Heuvel, "Annotation as a New Paradigm in Research Archiving", pp. 4–5.

oa:classifying seem most appropriate. It entails “the assignment of a classification type, typically from a controlled vocabulary, to the target resource(s)”. As is the case for the nanopublications data model, the OAC provides a mechanism for the description of the provenance of annotations. Such statements can be “useful for determining the trustworthiness of the Annotation, potentially based on reputation models”. Both the person and the time at which an annotation was created can be recorded. Provenance information can be attached to the Annotation, to the Body, and to the Target.

There are a number of important differences between the nanopublications framework and the OAC. One important difference is that, with the latter framework, it is possible to further modify the body of the annotation. The example below illustrates the manner in which the rhyme scheme of a stanza may be captured via OAC.[506] As can be seen, the Body of the annotation consists of the semantic tag lit:PerfectRhyme. This tag is classified as an oa:SemanticTag, and, additionally, it is associated with a pattern which represents the rhyme scheme. Next to the classes from the OAC and the ontology of literary terms, the code below also makes use of terms from FOAF, which is an ontology which can be used to describe properties of persons and of the relations between persons.[507] The person who takes responsibility for the annotation is identified using the oa:annotatedBy property.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix lit: <http://www.bookandbyte.org/olt/0.1/#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix oa: <http://www.w3.org/ns/oa#> .

<http://www.bookandbyte.org/macneice/annotations/anno1>
   a oa:Annotation ;
   oa:annotatedBy <info:eu-repo/dai/nl/32959737X> ;
   oa:hasTarget <AContact.xml#s2> ;
   oa:hasBody lit:PerfectRhyme ;
   oa:motivatedBy oa:classifying .

<lit:PerfectRhyme>
   a oa:SemanticTag, skos:Concept .
   lit:hasPattern “1 - 1” .
```

[506] Advice on how to express the data that were produced in this research project via OAC has kindly been given by Rob Sanderson and by Tim Cole.

[507] <http://www.foaf-project.org/> (26 April 2014)

```
oa:classifying
   a oa:Motivation .

<info:eu-repo/dai/nl/32959737X>
   foaf:name “Peter Verhaar” .
```

A second important difference between OAC and nanopublications is that, within OAC, it is also possible to create annotations which consist of multiple Targets. An occurrence of internal rhyme may be described using the following structure.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix lit: <http://www.bookandbyte.org/olt/0.1/#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix oa: <http://www.w3.org/ns/oa#> .

<http://www.bookandbyte.org/macneice/annotations/anno1>
   a oa:Annotation ;
   oa:annotatedBy <info:eu-repo/dai/nl/32959737X> ;
   oa:hasTarget <AutumnJournal-II.xml#w47-4> ;
   oa:hasTarget <AutumnJournal-II.xml#w48-2> ;
   oa:hasBody lit:AichillRhyme ;
   oa:motivatedBy oa:classifying .
```

One important consequence of the fact that OAC closely follows the central principles of the semantic web is that it can ultimately become possible to share annotations across applications and across technical platforms. Van den Heuvel and Roorda also identify a number of important challenges. At present, there is no widespread consensus concerning the manner in which text fragments may be targeted. Projects have often implemented idiosyncratic solutions, and this potentially complicates an exchange and a reuse of annotations. Secondly, since the OAC model does not prescribe a terminology that can be used within the body of an annotation, assertions created within different projects may also contain distinct vocabularies, undermining their interoperability.[508] The use of semantic web technologies in general equally implies a range of difficulties. Veltman notes that the semantic web “deals with meaning in a very restricted sense and offers static solutions”.[509] The technologies may be relevant for relatively straightforward fact-

[508] Dirk Roorda & Charles van den Heuvel, “Annotation as a New Paradigm in Research Archiving”, p. 4.

[509] Kim H. Veltman, “Towards a Semantic Web for Culture”, in: *Journal of digital information*, 4:4 (2006), p. 2.

finding purposes in scientific applications, but the humanities generally deal with less formalised information. Within humanities research, it is essential to take historical and geographical dimensions into account, as the meaning of terms may change both synchronically and diachronically. Meroño-Peñuela explains similarly that the sources which are maintained by cultural heritage institutions are often "messy and heterogeneous", and that such complexity can strongly complicate longitudinal queries. Humanistic ontologies, for this reason, demand "dynamic concept formalizations instead of static ones, especially for contested, open-textured or ambiguous concepts".[510] In many cases, it is necessary to capture data about the historical and the social contexts in which terms are used, and to allow for multiple, potentially contradictory conceptualisations of terms.

Since annotations based on semantic web technology often make use of pre-defined ontologies, scholars who use such languages need to describe the literary phenomena that occur in a text via fixed categories. Semantic enhancement generally implies a form of simplification, as the limited list of terms which were anticipated in an ontology does not necessarily match the diversity of the actual phenomena which are encountered in the literary text. A central difficulty that inheres almost inevitably in all classification systems is that there is a conflict between the demands that are posed by the need to process texts systematically on the one hand and the need to describe texts flexibly and responsively on the other. When a particular collection of texts illuminates the inadequacy of an existing vocabulary, scholars may respond to this by creating a new ontology, but the use of idiosyncratic terms may jeopardise the ability to query text collections consistently.

Andy Clark's thesis of the extended mind states that the technologies that we use to capture and to process information can take over crucial functions of the human brain. The semantic web may be viewed as one of the technologies which scholars can use to extend their minds and to stimulate and to enhance their thinking. If it is accepted that particular technologies also result in particular types of thinking, it may be expected that the processing based on semantic web technologies inhibits the development of radically new ideas. It places phenomena in categories which have been envisaged beforehand, and plaintively disregards the phenomena for which such categories do not apply. These difficulties, which can arise from the rigidity of ontologies, have partly been addressed within the Pliny tool, which was developed by Bradley and Prusin. Pliny can be used "in the pre-ontological context".[511] The tool provides support for the creation of annotations at the moment when they are still largely unstructured. Pliny firstly lets its users record any ideas that occur during the reading of the text that is being studied, in a

[510] Albert Meroño-Peñuela, "Semantic web for the humanities", in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, (Springer, 2013), p. 646.

[511] J. Bradley, "Thinking about Interpretation: Pliny and Scholarship in the Humanities", in: *Literary and Linguistic Computing*, 23:3 (5 September 2008), p. 271.

format that is not governed by an existing ontology. Once captured, users can rearrange and reshuffle the notes in order to "discover previously unrecognized patterns and relationships and to stimulate new ideas",[512] thus inspiring new interpretations. Pliny aims to provide a solution to the difficulty that generic preconceived ontologies rarely do justice to the particularities of an individual work of literature. Personalised annotations can be valuable for the interpretation of separate texts, but quantitative research performed on large corpora of texts typically demands that all the phenomena which are similar or comparable can also be referred to consistently using a stable set of descriptive terms. The conflict between generic descriptive terms and terms which are geared towards the particularities of a work of literature continues to pose an important conundrum.

Despite the challenges that can potentially complicate an exchange and a reuse of the data, The OAC data model appears to meet most of the requirements that were discussed in the introduction of this chapter. It can be used in combination with an ontology which supplies descriptive terms for the phenomena which have been discussed in Chapter 6 of this thesis. The data format in itself does not pose any technical difficulties. The targets of the annotations were formed by fragment URIs, and these are based on identifiers which are declared in the TEI files. The locations that are delineated by these identifiers can overlap, and the problem of conflicting hierarchies does not pose itself. Importantly, it is possible to express disagreement, as statements can be attributed exclusively to a single scholar. In cases of controversy, scholars who disagree may create separate annotations. A decision has been taken, for these reason, to record all the secondary data of the case study via the OAC.

7.4. Conclusion

This chapter has compared two types of data formats on the basis of three criteria. Formats were compared by considering the expressiveness of the ontology they are based on. Secondly, the analysis focused on the technical constraints that were connected to the syntax of the format. Thirdly, the possibilities for expressing disagreement were also taken into account.

With respect to the TEI, it was found that the default ontology it is based on[513] offers limited possibilities for the characterisation of aspects of poetic language. The standard, by default, does offer extensive support for the description of the logical structure of texts, for the identification of place names, geographic terms, bibliographic titles and for the description of the various witnesses of historical

[512] J. Bradley, "Thinking about Interpretation: Pliny and Scholarship in the Humanities", p. 266.

[513] In this context, the phrase "basic ontology" essentially refers to the collection of elements and the relationships between these elements that have been defined in the canonical tei_all schema. This core set of elements has been extended within specific research projects, but it is understood that such customisations do not belong to the standard's basic ontology.

texts. The standard is consequently very suitable in the context of textual criticism. It is less effective in studies which aim to explore literary phenomena across different texts on the basis of a quantitative method. The canonical TEI schema can obviously be modified, and those terms which were found missing can be added in a customised schema. Although such modifications may solve the problem of limited expressivity, the technical format of inline XML also poses a number of complications. Annotations generally target different textual fragments, and some of these targets clash. The strict hierarchical structure of XML prohibits a flexible delineation of such overlapping text fragments. Additionally, it is mostly impossible to record contrastive descriptions of text fragments within a single TEI document.

It was also shown that, when structural annotations are recorded using Semantic Web technologies, some of these complications can be avoided. Statements in RDF can derive their descriptive terms from an ontology, and when the terms which are necessary are not supplied by an existing ontology, a new conceptualisation of a domain may be supplied by making use of OWL. In this study, a new ontology was proposed for the description of literary phenomena. In a sense, an OWL-based ontology does not differ in a qualitative sense from a customised TEI schema. Both types of ontologies can, after they have been developed and tested within individual research projects, be made publicly available, and, after their publication, they may or may not attract a broader community of users. In the case of annotations stored via the OAC, however, most of the practical challenges associated with the TEI can be avoided. The targets of annotations are essentially references to specific locations within texts, and, if necessary, these locations may also overlap. The problem of conflicting hierarchies, consequently, does not present itself. Secondly, a single fragment can be described in many different annotations, and these annotations can all be associated with different scholars. The OAC can consequently be used to express polyvocality in literary criticism.

It may be observed that, within digital humanities studies, there is often a bifurcated position towards ontologies. In computer-based research, scholars typically use the descriptive terms supplied by a standard to describe their individual response to a literary text. The descriptions that are created are often time and location specific. They are often coloured heavily by the tenets of specific schools of criticism. At the same time, the ontologies which supply descriptive terms are typically perspectival as well. They are invariably developed for a specific purpose and within a critical tradition. As the form of usage which the standard accommodates does not necessarily match the needs of individual scholars, there is frequently an urge to manipulate the central ontology. The TEI standard can clearly be made responsive to idiosyncratic scholarly propensities. In large scale quantitative analyses of texts, however, it must be ensured that all instances of the phenomena which are studied are consistently identified as such. Quantitative research crucially requires the consistent application of a fixed ontology. In digital humanities research, it is often pivotal to strike a delicate balance between rigidity and extensibility.

An important characteristic of OWL is that it enables scholars to define relations between different ontologies. By making use of the property “sameAs”, it can be recorded, for instance that a term such as “lineGroup” from one conceptualisation is semantically equivalent to a concept which is referred to elsewhere as “stanza”. Similarly, the property “differentFrom” can be used to explain the disparities between ontologies. The aim to develop a single encompassing ontology seems a utopian quest. By making use of formal semantics, however, materials that have been formatted according to dissimilar methodological practices can still be integrated. When interconnections among distributed data collections have been made explicit via semantic web technologies, descriptions of distinct domains, or distinct descriptions of a single domain may still be reconciled.

In this study, a decision was taken to combine the TEI and the OAC framework to use the strengths of each. The interpretative observations have largely been captured via OAC, as it was estimated that the protocol allows both for strictness and for a degree of flexibility. This chapter has focused mostly on the representation of structured annotations. The data that are captured, however, obviously serve a specific purpose. They are created in order to perform specific analyses. The more precise ways in which the data can be analysed will be discussed in the next chapter.

Chapter 8

Data analysis

8.1. Introduction

Many of Louis MacNeice's poems express a fascination with the notion that changes in perspective can produce entirely new experiences. In "Under the Mountain", for instance, the poet captures the idea that the view from a mountain can miraculously purge a landscape of its everyday banalities. At ground level, "[t]he breakers are cold scum and the wrack / Sizzles with stinking life", but the mountain view reveals the poetry of the landscape and changes the bay into "a goose-quil that feathers ... unfeathers itself" (ll. 2-3). Similarly, in the early poem "Morning Sun", MacNeice concentrates on the transformative powers of sunlight. While daylight "blazons / The red butcher's and strolls of fish on marble slabs" (ll. 14-15), this "hanging meat / And tiers of fish are colourless and merely dead" at dusk (ll. 19-20). The poem "Mutations" proposes additionally that such unexpected vistas and startling perspectives are essential for the intellectual development of human beings. Shifts in perspective have the effect that established views may "crack at times" and that "new / Patterns from new disorders open like a rose" (ll. 8-9).

The urge to apply computing in the context of literary research is often driven by the analogous conviction that changes in perspective can foster new kinds of insights. In literary informatics, the changed vista generally involves a transition from a detailed examination of a limited number of texts to the possibility to focus on collections in their entirety. Texts in which concepts and literary techniques are originally presented linearly are transformed into a collection of discrete data, allowing for structural or synchronic analyses of full corpora. [514] When data are generated out of primary sources, and when the resultant annotations are stored separately from the original documents, it may be argued that these structured annotations collectively form a new and more encompassing resource. Within the data set that is created, the original division into books and volumes becomes less

[514] Writing in 1978, Susan Wittig conjectured that progress in the field of computer-based textual analysis was hampered by the fact that texts were habitually viewed as linear structures. This view, which, according to Wittig, was inherited from New Criticism, undervalued the non-linear or network-like patterns that can be produced by the phonological, syntactic or semantic units within literary texts. See Susan Wittig, "The Computer and the Concept of Text", in: *Computers and the Humanities*, 11 (1978), p. 212. Although the view on whether or not the original text should principally be viewed as a linear entity may differ, the results of algorithms for the creation of structured annotations about texts are typically discrete and context-insensitive.

consequential. In a sense, such large databases implement the logic that was envisaged for Vanevar Bush's memex. Bush proposed that knowledge which is traditionally organised according to physical boundaries of the printed work should be reorganised on the basis of the principle of "associative indexing, the basic idea of which is a provision whereby any item may be caused at will to select immediately and automatically another". In the memex, individual pages could also be "gathered together to form a new book". [515] The discrete data sets that are generated out of linear texts can likewise serve as proxies in which fragments from one particular text can be connected directly to fragments from other texts, purely on the basis of semantic or linguistic similarities. Gooding argues, more strongly, that digitisation projects decontextualize and deconstruct texts, to such an extent that there no longer are any books. The original texts are replaced by a vast corpus of words and metadata, which can all be divorced from the original context and which can collectively be searched uniformly.[516]

The benefits of digital tools often seem proportional to the volume of the text collection that is analysed. Tasks which require few data as input are generally performed most effectively by human beings, especially when such work demands interpretation. The analysis of corpora that span hundreds or thousands of works is strongly complicated, however, by the fallibility of human memory and the inconsistency of human concentration levels. Methodical comparisons of thousands of texts can generally be performed exclusively when scholars make use of digital research methods. In his monograph *Macroanalysis*, Matthew Jockers clarifies the value of computational analyses by drawing an analogy with the field of economics. Whereas microeconomics concentrates on individual companies or on individual families, macroeconomics aims to explain the economic behaviour of countries or of continents in their entirety.[517] Literary informatics comparably aims to uncover the larger trends in text collections via rigorous manipulation of large-scale aggregations of digital data. Other scholars have proposed similar metaphors to clarify the rationale of the shift to a larger scale. John Burrows compares large text corpora to handwoven rugs in which the "principal point of interest is neither a single stitch, a single thread, nor even a single color but the overall effect".[518] In the first pamphlet of the literary lab, Allison et al. use architecture as an image, suggesting that literary texts "like buildings, possess distinctive features at every possible scale of analysis".[519]

[515] Vannevar Bush, "As We May Think", in: *The Atlantic*, (1945), n.pag.

[516] P. Gooding, M. Terras & C. Warwick, "The Myth of the New: Mass Digitization, Distant Reading, and the Future of the Book", in: *Literary and Linguistic Computing*, 28:4 (13 August 2013), p. 428.

[517] Matthew Jockers, *Macroanalysis : Digital Methods and Literary History*.

[518] John Burrows, "Textual Analysis", p. 324.

[519] Sarah Allison et al., *Quantitative Formalism: An Experiment*, p. 8.

In this thesis' case study, data have been produced about a number of literary phenomena, and these structured data allow for a wide range of statistical analyses. This chapter discusses a number of ways in which the data can be processed.

8.2. Data analysis

As a rule, computer-based literary research is based on the central assumption that texts can be compared on a quantitative basis. While this reduction to numbers evidently implies a full neglect of the numerous unfathomable and unquantifiable ways in which poetry can produce effects, ensuring the statistical soundness of numerical comparisons poses challenges as well. The software that was developed for the purpose of this study has considered MacNeice's verse line by line. In agreement with Martin Mueller's observation that text analysis tools essentially produce lists,[520] the direct output of the application consisted of a simple index of all individual occurrences of the literary devices under investigation. To shift away from a focus on separate occurrences, and to be able to extend the scope to larger units, such as poems or volumes in their entirety, it is necessary to aggregate these individual values.

In most cases, instances of literary phenomena can be counted in a variety of ways. In the case of alliteration, for instance, metrics can be produced by counting the different repeated sounds that occur within a line. Alternatively, metrics may also reflect the number of words that alliterate, or the number of verse lines within a poem that actually contain alliteration. Applying these three distinct methods to the line "This wind from the west be backed by waves" (l. 94), in "Country Weekend", for instance, would result in values two, six and one respectively. The number of rhyming sounds in a poem can similarly be counted in different ways. Stanza totals may either reflect the number of repeated sounds within a stanza or the total number of lines that rhyme. The second stanza of the poem "Passage Steamer", for instance, consists of 7 lines which end in "require", "desire", "sun", "none", "mast", "past" and "begun", making the rhyming scheme AABBCCB. In this situation, there are obviously seven lines that rhyme, but only three different rhyming sounds. In this situation, the choice of the summation method depends on the objective of the analysis. Counts of the lines that rhyme give an impression of the overall regularity of the stanza. Counts of rhyming sounds, by contrast, indicate either the monotony or the versatility of the lines that rhyme.

Counts are rarely neutral, as they invariably demand decisions on what and how to count. Such discussions of the various methods for producing counts are not trivial, as decisions on how to count can directly have a strong impact on the outcomes of subsequent statistical processing. As a result of the relative novelty of this type of research, few best practices have emerged. Transparency on how

[520] Martin Mueller, "Digital Shakespeare, or towards a Literary Informatics", p. 291.

numbers have been produced, therefore, is of crucial importance in literary informatics. In the case of alliteration, it was observed that there are important differences in the intensity of occurrences. Lines such as "Further who failed last Friday to feel grieved" in "Easter Returns" (l. 11), or "The flotsam of private property pekinese and polyanthus" in "An Eclogue for Christmas" (l. 100) clearly alliterate heavily. These examples differ in a qualitative manner from a line such as "Round the Corner is sooner or later the sea" (l. 2) in "Round the Corner", as this line only contains one repetition of an /s/ sound. In this study, it was decided to give expression to the intensity of occurrences of alliteration by counting the number of words that alliterate. The same decision has been taken for occurrences of assonance and consonance. With respect to rhyming sounds, it was decided to count the number of lines that rhyme, rather than the number of rhyming words, as analyses of rhyme are mostly aimed at determining the regularity of stanzas.

An additional complexity is produced by the fact that the poems in the corpus obviously differ in length. When different poems need to be compared directly, it does not seem legitimate to compare texts on the basis of absolute counts. In a long poem such as "Valediction", which contains 105 lines and 856 words in total, the author obviously has much more opportunities to use alliterative language than in "Aubade", which only contains 53 words. It has been decided, for this reason, to correct all absolute numbers, by dividing the total number of occurrences either by the total number of tokens or by the total number of lines. Alliteration is obviously a device which is applied at the level of tokens. In theory, an author can introduce alliterations with each new word that is used, and the total number of alliterative sounds can never be higher than the total number of tokens. For this reason, alliteration, assonance and consonance have been normalised using the token counts. Devices such as perfect rhyme, slant rhyme and semi-rhyme, however, operate at the level of the verse line, and the absolute counts of all occurrences or rhymes have been normalised on the basis of the total number of lines. On the basis of these principles, normalised values have been produced for all the 311 poems in the corpus. The long poem "An Eclogue for Christmas" contains 1444 tokens, and the absolute counts for alliteration, for instance, amounts to 253. The normalised value is (253 / 1444 =) 0.175. The shorter poem "Invocation", which has 146 words, has 52 alliterative words. The relatively high intensity of alliteration in this poem is expressed in the normalised value 0.356. During initial tests, an additional measure proved to be necessary to ensure the accuracy of the normalisation method for counts of rhymes. A number of poems contain stanzas which consist of one line only. One example is MacNeice's "Autobiography", in which the haunting refrain "Come Back Early or Never Come", is not contained within a longer stanza. When stanzas contain one line only, it is evidently impossible for these to contain rhyme. The number of single-line stanzas have therefore been deducted from the total number of lines before the normalisation.

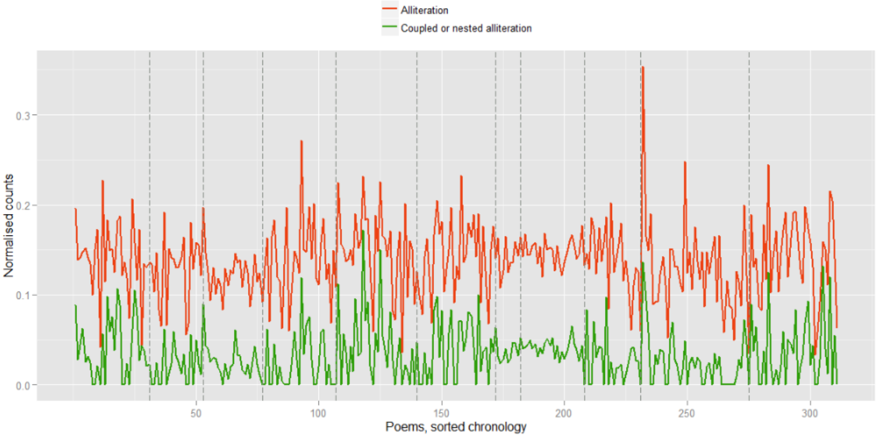


Figure 8.1. Normalised counts of occurrences of alliteration

The values that have been calculated can be used, firstly, to trace particular historical developments at a macro-level.[521] Figure 8.1 displays the distribution of the counts for alliteration across MacNeice's entire oeuvre. In the diagram, the poems have been sorted according to their location in the *Collected Poems*, edited by Peter MacDonald, and published in 2007. This arrangement roughly corresponds to a chronological order. The sections which are distinguished using the dashed vertical lines correspond to the 11 volumes in which these poems have appeared. As can be seen from the peaks and the troughs in the diagram, the values that have been calculated for alliteration fluctuate heavily during most of MacNeice's career. The usage is relatively stable, nevertheless, in *Autumn Journal*, *Ten Burnt Offerings* and *Autumn Sequel*.[522] As was also explained in Chapter 6, MacNeice frequently composed verse lines containing notable combinations of alliterative words. Many lines contain alliterations nested within another alliteration (e.g. "As gay trams run on tracks and cows give milk" in line 16 of "An April Manifesto" or "Greek words sprout out in tin on sallow walls" in line 324 of "Eclogue from Iceland") or structures in which two sounds in the first half of the line are repeated in the second half (e.g. "The ancient smiles of men cut out with scissors and kept" in line 17 of "Perseus" or "Columns of ads the quickest roads to riches" in line 25 of "Christmas Shopping"). The green line in figure 8.1 represents all occurrences of such specific forms of alliteration. This analysis indicates that MacNeice used such nested or alternating structures in most of his poems, and,

[521] The visualisations that are discussed in this chapter were created in R or in Processing. The code that was used to create these visualisations can all be found at <https://github.com/peterverhaar/Phd>.

[522] In figure 8.1, the poems in *Autumn Journal*, *Ten Burnt Offerings* and *Autumn Sequel* have been assigned the numbers 53 -77, 172 -182 and 183-208 respectively.

additionally, that the ratio between regular alliteration and such specific forms of alliteration is more or less stable.

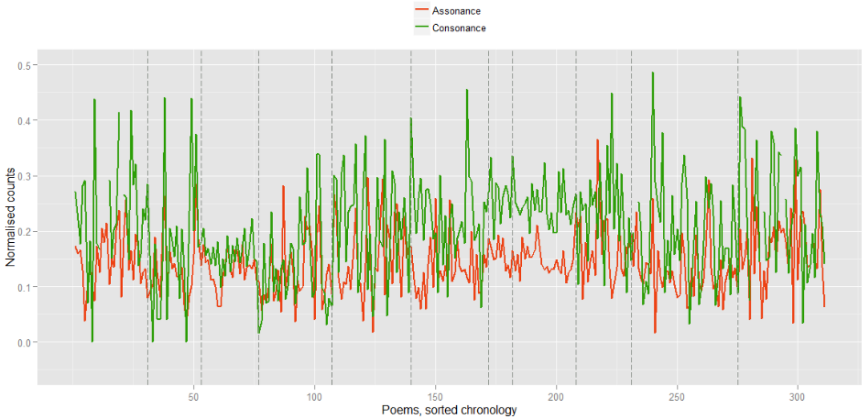


Figure 8.2. Normalised counts of occurrences of assonance and consonance

Figure 8.2 shows that MacNeice likewise made a relatively solid use of consonance and assonance. In general, the counts for consonance are higher than those for assonance. As is the case for the use of alliteration, the poems in *Autumn Journal*, *Ten Burnt Offerings* and *Autumn Sequel* display less variability with respect to the occurrences of assonance and consonance. Figure 8.2 also reveals, interestingly, that there are three poems which are fully without any occurrences of consonance. In this study, consonance has been implemented strictly as “the repetition of the sound of a final consonant or consonant cluster in stresses, unrhymed syllables near enough to be heard together”.[523] Although some authors view alliteration as a specific form of consonance in which sounds are repeated at the beginning of stressed syllables, this study has distinguished alliteration and consonance more sharply. The three poems which were found to be without consonance are “The Sunlight on the Garden”, “Cuckoo” and “Sand in the Air”. Consonance, like assonance, can be used to suggest connections between words, and, fittingly, the poems which are devoid of consonance all describe a loneliness or an isolation. “Cuckoo” references Symeon the Stylite who was solitary on his pillar and who formed the “Personification / Of distance” (ll. 7-8). “Sand in the Air” and “The Sunlight on the Garden” both describe a feeling of intense desolation after being abandoned by another person. The persona in the former poem bemoans his apprehension that in a “shrivelled world / There is only I” (ll. 21-22). The fact that the three poems that were mentioned lack consonance can easily be

[523] “Alliteration”, in *The Princeton Encyclopedia of Poetry and Poetics*, p. 299.

overlooked in studies based on traditional methods. Human scholars tend to concentrate on the textual phenomena which are present, and the basic fact that certain other phenomena are absent is often difficult to notice.

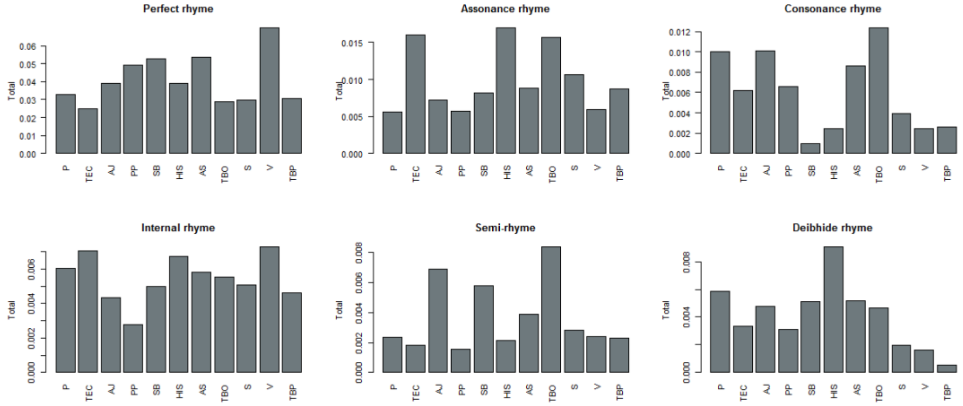


Figure 8.3. Normalised counts for perfect rhyme, assonance rhyme, consonance rhyme, internal rhyme, semi-rhyme and deibhide rhyme, aggregated at the level of volumes

To allow for a more focused comparison of the use of literary techniques throughout MacNeice's career, the normalised metrics can also be aggregated at the level of volumes.[524] Figure 8.3 displays the main differences in the use of the various forms of rhyme. The diagram was created, firstly, by counting all the occurrences in the various volumes, and, secondly, by dividing these counts by the total number of tokens in each volume. As can be seen, perfect rhyme has been used most frequently in MacNeice's penultimate volume *Visitations*, and it has been used most sparsely in the second volume *The Earth Compels*. For some volumes, the frequencies of perfect rhyme are inversely proportional to the frequencies of slant rhyme. *Visitations* has low scores for both assonance rhyme and consonance rhyme, while *The Earth Compels* has one of the highest values for assonance rhyme. The diagram also reveals that the volumes *Springboard* and *Holes in the Sky* appear to make use of a very specific form of slant rhyme. In these volumes, there are many occurrences of assonance rhyme, and relatively few occurrences of consonance rhyme. Other interesting findings are that the use of semi-rhyme is most frequent in *Autumn Journal* and *Ten Burnt Offerings*, and that the use of internal rhyme roughly remains at the same level throughout the

[524] *Autumn Journal* and *Autumn Sequel* are evidently not volumes. They are lengthy poems consisting of separate cantos. For ease of reference, however, this text will refer to these collections as volumes as well.

poet’s career, with the exceptions of *Autumn Journal* and *Plant and Phantom*. In general, the frequencies of internal rhyme, semi-rhyme and deibhide rhyme are much lower than those of perfect rhyme and slant rhyme. The use of deibhide rhyme is relatively stable in MacNeice’s first eight volumes, with a very sudden upsurge, however, in *Holes in the Sky*. The usage of this Celtic rhyming device diminishes quite abruptly in the last three volumes.

In this case study, data have also been collected about the imagery that occurs in MacNeice’s poems. Figure 8.4 clarifies a number of historical developments in the use of imagery. The diagram was produced by dividing the absolute number of references in each volume by the total number of tokens in each volume. The sizes of the circles reflect the number of references to the various images.[525] On the basis of the normalised counts of all references, it can be observed that the images “Religion”, “Plants”, “Transportation”, “Light” and “Water” have been used most profusely within the corpus. The counts provide support for Brown’s observation that references to the sea and to trains and other modes of transportation are frequent in MacNeice’s verse. The volume entitled *Plant and Phantom* also contains the highest number of references to plants. Religious imagery is used consistently throughout the entire career, and most frequently in *Springboard*, *Holes in the Sky* and *Visitations*. In the early volumes *Poems* and *The Earth Compels*, there are mainly reference to Christian religion, and the passages in which they are used often stress a conflict between religious ideals and the observed practices of a secular society. In “An Eclogue for Christmas”, people decorate their houses with “tinsel and frills” to “announce that Christ is born among the barbarous hills” (ll. 8-9), and people principally worship “the cult of every technical excellence” (l. 153). “Belfast” portrays a beggar woman “[t]o whose prayer for oblivion answers no Madonna” (l. 20), and “Birmingham” describes the failed “endeavour to find God” (l. 15). The volume *Plant and Phantom* refers more frequently to non-Christian religions and to mythology. “Stylite”, for instance, describes a “white Greek god” (l. 21), “Plant and Phantom” mentions the “pawky Fates” (l. 9), and “Novellettes” refers to an “Aztec pyramid of sacrifice” (l. 2017).

Whereas MacNeice wrote in the *Poetry Book Society Bulletin* that his volume *Solstices* “contains practically no allusions to either Graeco-Roman or Christian legend”,[526] this study did identify many references to Christian religion in this volume. “Bad Dream”, for instance, contains references to a bible and a “crucifix on the wall” (l. 22) and “Jericho” mentions “the fires of Pentecost” (l. 24). “Il Piccolo Rifiuto” describes the irritability that can result from the sense of being lost in translation in another country. The persona noticed that, during moments of

[525] During various trials and errors, I discovered that more common visualisation techniques, such as line diagrams or bar charts could not effectively be used to represent the developments in the use of imagery. The visualisation in figure 8.4 was programmed using the Processing environment.

[526] Louis MacNeice, Peter McDonald (ed.), *Collected Poems* (London: Faber & Faber 2007), p. 795.

miscommunication, “God / began to limp” (ll. 12-13). “The Blasphemies” and “The Messiah” probably contain the most obvious references to Christian religion. The former poem describes the persona’s troubled fidelity to religious faith. As a young boy, the persona is afraid to commit the “sin against the Holy” (l. 1), while in later life, he “turned to parody / Prayers, hymns, the Apostles' Creed” (ll. 14-15). “Windowscape”, furthermore, is a poem which focuses on a lacklustre suburb, which reinforces its central sense of doom and fatalism by stressing the obsolescence of religious practices. Although the “[w]indow-cleaner and postman call just once a year”, there is “never a priest” (ll. 8-9).

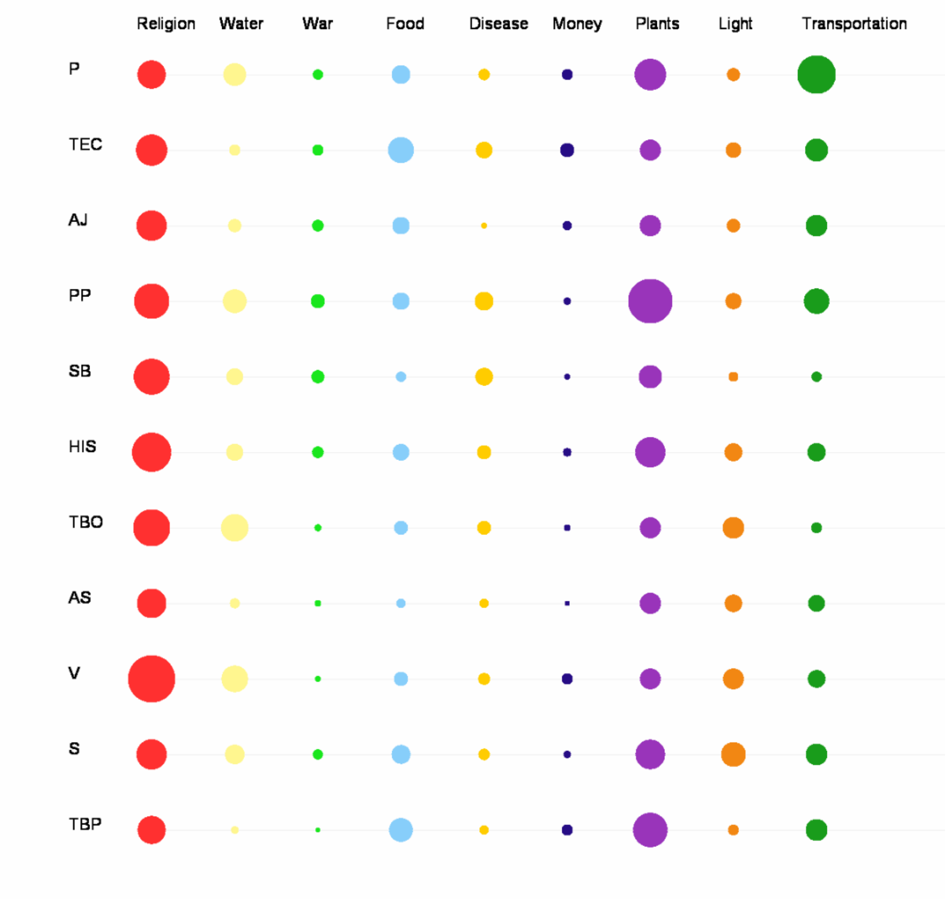


Figure 8.4. Use of imagery

One notable finding is also that war imagery remains relatively stable throughout the entire oeuvre. The war imagery is not significantly higher in the volumes written during the Second World War. In *Poems* and *The Earth Compels*, war imagery is often employed to evoke a general sense of doom, or to underscore the menacing aspect of forces that represent change. In “Spring Voices”, the coming of spring is presented as a preparation for a war. The new season is “massing forces”, and the main character is afraid “of the barrage of sun that shouts cheerily” (l. 2). In “June Thunder”, the sudden storm is presented as “the sword of the mad archangel” (l. 23). Within the poem “Brother Fire”, which is arguably the poem in which MacNeice addresses the atrocities of war most poignantly, no explicit war images were found.

Throughout MacNeice’s corpus, copious references were also found to consumer articles and to objects from popular culture. Images such as “Food”, “Drinks”, “Media”, “Sports” and “Money” all have comparatively high frequencies. References to money and to wealth often occur, expectedly, in passages which express a criticism of a capitalist and a consumerist society. The poem “An April Manifesto”, for instance, demands that “April must replenish / Our bank-account of vanity” (l. 10). Similarly, in “Christmas Shopping”, the shop windows “marshal their troops for assault on the purse” (l. 13). Interestingly, however, words connoting financial opulence are also used frequently to evoke a sense of happiness or of vitality. In “Snow”, for instance, the persona finds that the room “was suddenly rich” (l. 1). When the first shepherd in “Eclogue by a Five Barred Gate” recounts his blissful dream, he stresses that the light he saw was “delicate as the chink of coins” (l. 103). Additionally, MacNeice stresses in “Train to Dublin” that, by celebrating the transient phenomena which are enumerated towards the end of the poem, people can “find that they are rich and breathing gold” (l. 55). In the depiction of scenes from his marital life in *Autumn Journal*, moments of felicity are often associated with affluence. The persona recalls that “We slept in linen, we cooked with wine / We paid in cash” (“Autumn Journal VIII”, l. 38) and that “ till life did us part I loved her with paper money” (“Autumn Journal I”, l. 67). MacNeice's rejection of consumerism is apparent from poems such as “Christmas Shopping” and “In Lieu”, but the imagery that is used throughout his verse also prompts a reconsideration of the poet's position towards the pursuit of wealth. MacNeice's oeuvre, furthermore, contains many images of a gastronomic nature. In “Littoral”, for instance, the shore is compared to “[d]amson whipped with cream” (l.2). In “The National Gallery” the persona sees a “pink temple of icing-sugar” in a picture (l.12). The fragrant trees which were remembered from childhood in “When We Were Children” were “breakfast for the gluttonous eye” (l. 10). The poem “Constant” depicts the remnants of the various civilizations in Istanbul via a complaint that there are “[t]oo many curds on the meat” (l. 1). This cursory survey of imagery illustrates that MacNeice often manages to convey profound concepts effectively via objects taken from quotidian life.

Most critics agree that MacNeice's oeuvre can be divided into three broad phases.[527] The poetry written during the second phase in the early 1950s is mostly considered to be of a more languorous and a less inspired nature. The data set that was developed in this study can be used to investigate whether or not the changes in the critical acclaim also correlate with specific formal properties of the poetry. Figures 8.1 and 8.2 both suggest that specific clusters of poems are distinctive within the context of MacNeice's oeuvre not because of a particularly high or low value for a specific metric, but, rather, because of a lack of variation. As a lower variation might also imply a more predictable and a less creative use of devices, the hypothesis may be formulated that there is a correlation between the perceived quality of the poems and the ability of the poet to vary the application of literary techniques.

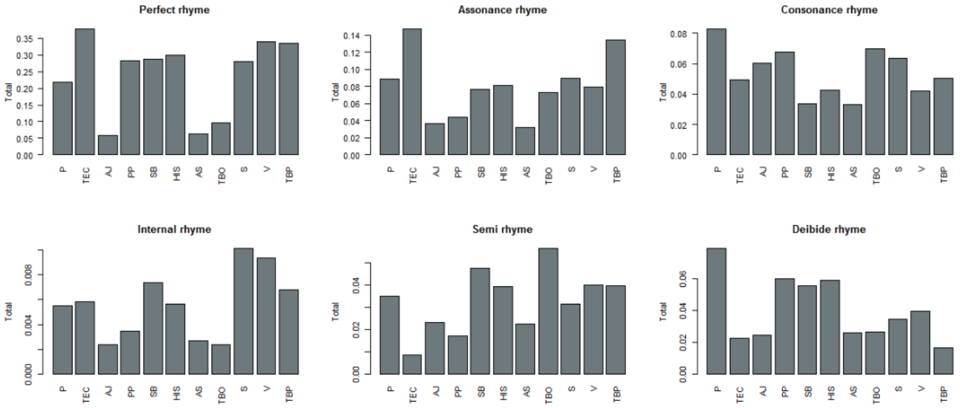


Figure 8.5. Standard deviations of volume-level counts of perfect rhyme, assonance rhyme, consonance rhyme, internal rhyme, semi-rhyme and deibhide rhyme

To explore the hypothesis more precisely, standard deviations have been calculated for a number of variables in all eleven volumes. [528] The results of these

[527] Alan Gilles explains that "[t]he contours of Louis MacNeice's career are rarely contested: from the high point of his nineteen thirties work, reaching a crescendo with *Autumn Journal* (1939), he drifted into a slump, reaching a nadir with two collections from the early nineteen fifties, *Ten Burnt Offerings* (1952) and *Autumn Sequel* (1954), before reviving to develop a startling new style at the end of the decade". See Alan Gillis, ""Any Dark Saying": Louis MacNeice in the Nineteen Fifties", p. 105.

[528] Standard deviations are calculated by taking the square root of the variance of all the values. The variance, in turn, is produced by calculating the average of the squared differences between all values and their mean. The standard deviation gives an indication of the variation within a collection of data values. When its value is close to zero, this implies that there is little variation. High values, conversely, indicate that values are more diverse.

calculations are shown in figure 8.5. Interestingly, it can be observed that *Ten Burnt Offerings* and *Autumn Sequel* largely have the lowest values of all volumes. While some other volumes combine low volumes for one variable with high values for other variables, the poetry of the early 1950s scores lowest on almost all metrics. An exploration of the variations in the occurrences of literary devices cannot fully account for the differences in critical acclaim, however. *Autumn Journal* is frequently regarded as one of MacNeice's best works, but the standard deviations that were calculated for this collection are largely on the same level as the poetry of the early 1950s. The stasis in the usage of literary devices can partly be explained through the fact that *Autumn Journal*, *Autumn Sequel* and *Ten Burnt Offerings* all contain long poems with relatively consistent rhyme schemes. *Autumn Sequel*, for instance, consistently uses the terzima rhyme form, in which the first and the third line of each triplet rhymes. For this reason, it must be concluded that the critical acclaim does not correlate completely with the variability in the use of literary devices.

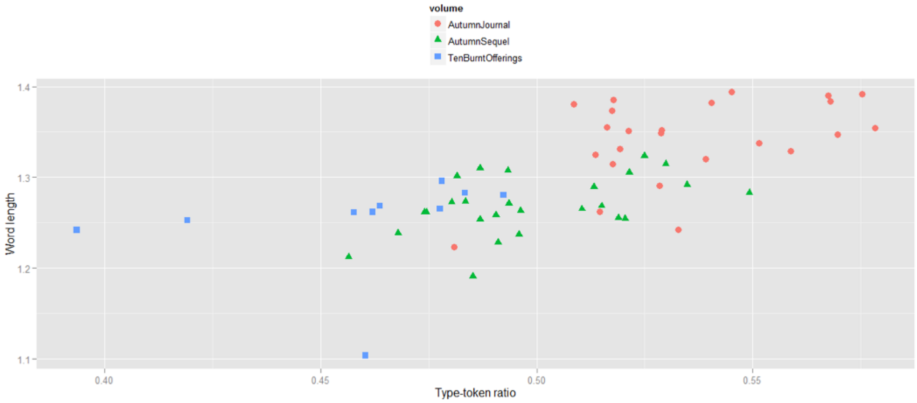


Figure 8.6. Type-token ratios and average number of syllables per word for all poems in Autumn Journal, Autumn Sequel and Ten Burnt Offerings

Prompted by the cursory impression that the language that is used in *Autumn Sequel* and in *Ten Bunt Offerings* is generally more rambling and more stagnant than that of *Autumn Journal*, I have created a scatter plot which shows both the type-token ratios and the average word lengths for all the poems in these three volumes. The word lengths were calculated by dividing the total number of syllables in each poem by the total number of tokens. Figure 8.6 indicates that the values that were calculated for *Autumn Journal* clearly differ from those of the two collections that were published in the early 1950s. On the whole, the poems in *Ten Burnt Offerings* and in *Autumn Sequel* have lower type-token ratios, which implies that there is also a lower diversity in the vocabulary. Additionally, the poems in the latter two collections poems also contain words with fewer syllables. *Autumn Sequel*, in fact, contains many monosyllabic lines, such as “The wits of Bath for all their sense of form” (“Canto XXI”, l. 105), “Put back no clock; clocks were made for men” (“Canto I”, l. 16) or “We are bound to live and give and make and act” (“Canto XVI”, l. 52). On the basis of these findings, it may be hypothesised that that this more repetitive verbiage, and the many reiterations of monosyllabic words, strongly contribute to the overall sense of aridity in the latter two collections.

The analyses that have been performed to this point largely concentrated on the differences and the similarities between the various collections of poems. As was also alluded to in the introduction to this chapter, however, computational methods also stimulate scholars to study the text in a corpus as nodes in an interconnected network, and to abate the original boundaries formed by the physical books and volumes which originally contained these texts. Writing about theories of literature, Wellek and Warren distinguish approaches which view literature as “a series of works arranged in a chronological order and as integral parts of the historical process”, on the one hand, and approaches which view literature as “a simultaneous order”, on the other.[529] In this latter approach, texts are juxtaposed on the basis of their stylistic features and regardless of when they were written. Computational methods support both approaches to the study of literary works. Using the data that were generated, poems can be grouped on the basis of a correspondence in formal properties, and the individual poems that are set apart in this manner can subsequently be examined in a more targeted manner.

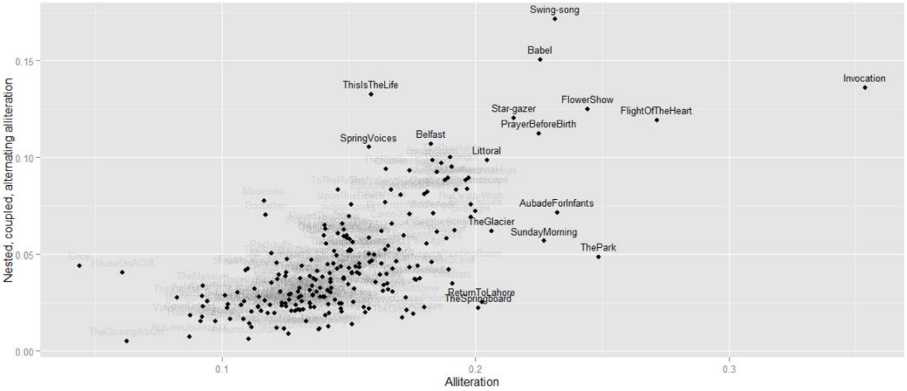


Figure 8.7. Normalised counts of alliteration in general and normalised counts of specific forms of alliteration

[529] Rene Wellek & Austen Warren, *Theory of Literature*, p. 39.

Figure 8.7 shows the normalised counts of all occurrences of alliteration for each poem, together with the counts of the structured types of alliteration, in which two pairs of matching sounds are combined in specific ways. Notably, some of the poems which have high scores for alliteration also have similar themes. “Invocation” and “Flower Show”, for instance, both describe a withdrawal from a physical reality, followed by an escape into a dream-like or imaginative world. The heavy alliterations in “Invocation” (e.g. “Fetch me far a moon in a tree / Fetch me far a phrase of the wind”, ll. 15-16) and in “Flower Show” (e.g. “Fanged or whaleboned wattled or balding brimstone or cold / As trout or seaweed these blooms ogling or baneful all”, ll. 8-9) have an incantatory effect which helps to evoke the otherworldly nature of the environment which is depicted. Alliteration is used in a comparable manner in “Littoral”. In this poem, the persona is swept away by the many associations that are provoked by the image of the sea. Two other poems which are highlighted in figure 8.7 are “Babel” and “Flight of the Heart”. Next to the fact that these poems both feature a tower, they are both about an impulse to move away from social and political realities. The former poem uses the myth of Babel to underscore the social conflicts that can exist between the denizens of a modern city, and between “[p]atriots, dreamers, die-hards, theoreticians” (l. 13). “Flight of the Heart”, which describes an urge to fully retreat into the cellar of a tower, may be read as a reflection on the relation between the self and the external world. Terence Brown stresses that the separation between the self and its physical environment is a “foregone conclusion” in “Flight of the Heart”. In “Babel”, however, the questions “Can't we ever, my love, speak in the same language?” and “Have we no aims in common?”, which are repeated insistently in each stanza also implies a hope that the distances can ultimately be bridged.

The historical overview of the occurrences of literary devices in figure 8.3 suggests, among other things, that semi-rhyme has not been used consistently throughout MacNeice’s career. It may be instructive, for this reason, to concentrate more closely on the poems that contain examples of this particular form of rhyme. Figure 8.8 displays the poems which, when the corpus is sorted according to the scores for semi-rhyme, belong to the upper five percent. Semi-rhyme evokes the impression of an order which is incomplete, or of a uniformity which emerges illogically from chaos. This device is often used in poems which depict an eerie atmosphere, or to evoke a confusing or a chaotic reality. “The Pale Panther” has one of the highest scores for this particular form of rhyme. In this poem, MacNeice connects “lamented” to “bent” and “cure” to “surely”. MacKinnon describes the poem as one of MacNeice’s nightmares, in which “the mood of bleak despair is not balanced by any of his old sardonic optimism”.[530] The semi-rhymes in “Bottleneck” and “Spring Voices” have similar effects. The former poem argues that social and

[530] William T. McKinnon, “MacNeice’s Pale Panther: An Exercise in Dream Logic”, in: *Essays in Criticism*, XXIII:4 (1973), p. 388.

political questions are often too complex or too confounding to be able to take a firm definitive stance. The main character's idealistic but confused politics aim to be "combined / Into a working whole but cannot jostle through / The permanent bottleneck of his highmindedness" (18-20). In "Spring Voices", the uneasy association of words such as "air" and "warily" (ll. 1-3) effectively reinforce the central opposition between the lures of the sunlight and the new spring on the one hand, and the nihilism and the apathy of the "householder" which is depicted on the other. Like "Spring Voices", "Nostalgia" describes a struggle between two impulses. The urge to focus on actuality and on facts can be countered by a pressing "longing / For what was never home" (ll. 3-4). This conflict is mirrored, to some extent, between the uncomfortable conjunction of the words "lull" and "vulnerable" and "slow" and "lonely".

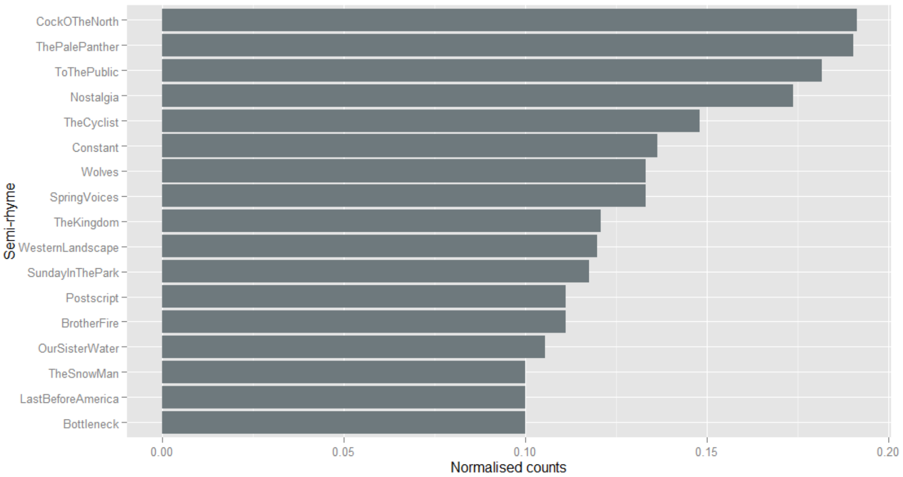


Figure 8.8. Poems with the highest normalised counts for semi-rhyme

A similar form of analysis can be performed to study regularities in the usage of deibhide rhyme. Figure 8.9 lists the individual poems which have the highest values for this specific device. The deibhide rhyme juxtaposes pairs of words which sound similar but which nevertheless do not agree fully. Reviewing the poems that are culled in figure 8.9, it may be concluded that such sonic agreements between words with dissimilar stress patterns have frequently have been used in poems which are concerned with social differences and with interpersonal conflicts. "Conversation" and "A Contact" form two clear examples. In the former poem, MacNeice proposes similarities between the words "way" and "yesterday" and "straight" and "interpolate". The poem concentrates on the uneasy moments during which the customary superficiality of social interactions is interrupted by inadvertent outbursts of genuine emotions. "A Contact", which rhymes "windows" with “goes”, muses on the differences between a train that is chosen and the trains that pass by, and uses this imagery to evoke a sense of isolation and separation.

The notion that MacNeice frequently used deibhide rhyme to emphasise a strained relation between the self and the other is also exemplified clearly by some of the other poems which are enumerated in figure 8.9. “The Individualist Speaks”, for instance, contains two examples of deibhide rhyme: “As chestnut candles turn to conkers so we / Knock our brains together extravagantly” (ll. 5-6) and “Crawling down like lava or termites / Nothing seduces nothing affrights” (ll. 13-14). The poem as a whole imparts MacNeice’s mistrust of established creeds and of existing philosophical systems. Ideologies such as socialism are depicted as exhibits in a fair “pitched among the feathery clover” (l. 1). The persona, however, ultimately chooses to “escape, with my dog, on the far side of the Fair” (l. 16). The use of deibhide rhyme appropriately underscores the poem’s tension between communality and individuality. Deibhide rhyme is also used in the first and third stanzas of “Birmingham”. In the description of the slums which houses the factory workers, the “vista thins like a diagram / There unvisited are Vulcan's forges who doesn't care a tinker's damn”. In the third stanza, the “shopgirls’ faces relax / Diaphanous as green glass empty as old almanacs”. Rhymes of stressed and unstressed syllables evoke a sense of inequality, and “Birmingham” productively uses this device to accentuate social distances. Along similar lines, “Bar-room matins” describes an inability or an unwillingness to empathise with the victims of World War II. It connects the words “sky” and “alibi” (ll. 22-23) and “sea” and “commodity” (ll. 26-27).

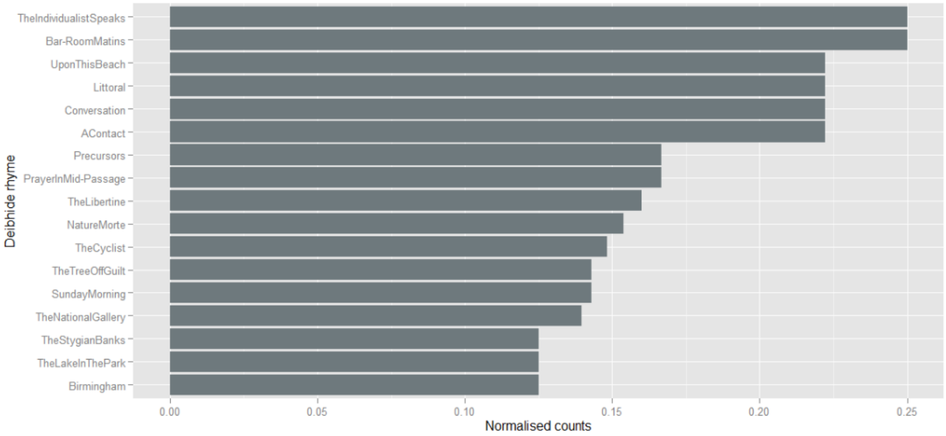


Figure 8.9. Poems with the highest normalised counts for deibhide rhyme

The list that is shown in figure 8.9 was produced by sorting the poems on the basis of their normalised counts. Figure 8.9 only lists 17 out of the 116 poems which contain instances of deibhide rhyme. During a closer inspection of the remaining poems, several other striking instances were found. It is interesting to observe, for instance, that this device, which was derived from the Celtic verse tradition, is also used in “The Hebrides” and in “Valediction”, which both describe parts of the British Isles with a strong Celtic heritage. [531] “Valediction” describes the city of Belfast as “devout and profane and hard / Built on reclaimed mud, hammers playing in the shipyard” (ll. 15-16) and warns all tourists in Ireland not to “pay for the trick beauty of a prism / In drug-dull fatalism” (ll. 65). The two poems express dissimilar messages about the regions they portray, however. “The Hebrides” idealises a traditional and rural way of living, and also hints at some of the negative implications of modernity and consumerism. “Valediction” does not idealise Ireland’s past, but emphasises the asphyxiating effects of Ireland’s obsession with its history and with its religion. Notable instances of deibhide rhyme can also be found in “Les Sylphides” from “Novelettes”, which rhymes “ballet” with “grey”, and in “The Libertine” which connects “alone” to “telephone”. The latter two poems provide additional support for the claim that deibhide rhyme is often used in passages which concentrate on interpersonal distances. “Les Sylphides” depicts a failed marriage, and “The Libertine” explores the bewilderment of a man who, after a life full of promiscuity, only desires to be left alone. In the context of algorithmic criticism, the objective of data analysis is generally to expose aspects of literary texts which can enkindle interpretation, and, in view of this aim, algorithms ought to be able to detect remarkable instances of literary phenomena. Given a particular data set, however, it can often be challenging to develop an effective method for the discovery of significant fragments. Since the ability to advance interpretation seems difficult to quantify, an approach which is based solely on the number of occurrences does not necessarily form a suitable heuristic for ensuing interpretative readings.

The comprehensive data set that was developed in this study can productively be actuated to expand the results of previous critical studies. In a discussion of *Leaving Barra*, Neil Corcoran notes that the poem consists almost exclusively of feminine line endings. Corcoran argues that the lowering intonation in these lines, in which a stressed syllable is followed by one or more unstressed syllables has a melancholy connotation, evoking “courage in the face of loss, something falling poignantly away”.[532] Using the data set that was developed in this study, it is possible to investigate if such lines endings have similar effects in other poems. Via

[531] Next to deibhide rhyme, these two poems also contain other examples of devices which, according to Skelton, were inherited from Celtic verse. “The Hebrides” contains both regular aicill rhyme (“Are glad to have their land though mainly stones / The honoured bones which still can hoist a body”, ll. 52-53) and a specific form of aicill rhyme in which the final consonants of a line are repeated in the line that follows (“And all the neighbours celebrate their wedding / With drink and pipes …”, ll. 84-85). A similar example can be found in lines 51 and 52 of “Valediction”: “Of green marble or black bog-oak run up to Clare / Climb the cliff in the postcard visit Galway city”.

[532] Neil Corcoran, “The Same Again? Repetition and Refrain in Louis MacNeice”, p. 215.

the phonetic transcriptions that were made available, poems with similar properties can be identified relatively swiftly. For each poem, the lines which end in unstressed syllables were counted, and, after this, these counts were divided by the total number of lines. Figure 8.10 displays the 20 poems with the highest percentages of feminine line endings. The yellow lines indicate feminine line endings, and the blue lines indicate masculine line ending. Most of the poems in this group of texts indeed focus on a loss. "June Thunder" concentrates on the poet's divorce from his wife, and the stream in the final stanza of "The Rest House" is read by Brown as "some allegorical river of life, winding from birth to death".[533] "New Jerusalem", furthermore, depicts the city of London, which changed from a place that could excite and inspire into a location full of "compulsive, illusory distractions born out of a society more firmly tethered to a rampant consumer culture".[534]

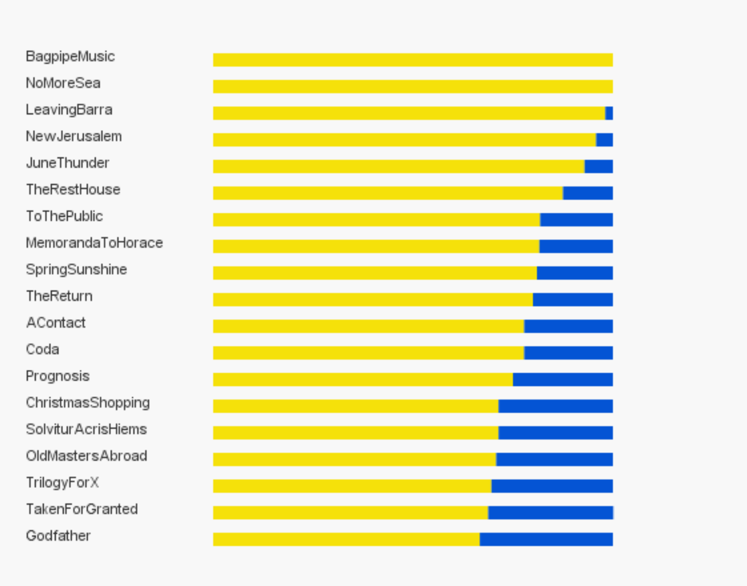


Figure 8.10. Proportions of feminine and masculine line endings

Interestingly, the three poems in the corpus whose final words are almost exclusively feminine all focus on islands. "Bagpipe Music" focuses on the decay of a harmonious rural culture on the Western Scottish islands, and on the transition to

[533] Terence Brown, *Louis MacNeice: Sceptical Vision*, p. 141.

[534] Simon Workman, ""To Be Tired of This Is to Be Tired of Life": Louis MacNeice's London", in: *Irish Writing London: Volume 2*, Tom Herron, London: Bloomsbury Academic, p. 139.

a consumerist and violent society. Brown notes that the poem has “a nihilistic cruelty in its pointless violence”.[535] “Leaving Barra”, as discussed by Corcoran, describes the poet’s continuing need for religious beliefs or for elucidatory theories. The departure from the island of Barra symbolises the concurrent inability to commit to such creeds. The thematic concerns of “No More Sea” are strongly related to those of “Leaving Barra”. In the latter poem, the sea represents a “vast nothingness”.[536] Human beings can shield themselves against such nihilism either by devising individual guidelines, or by joining the accepted rules and dogmas of a larger group. “No More Sea” represents the clash between individuality and communality as the difference between islands and the mainland, and advocates the value of islands. In all these three poems, the islands represent a moral or spiritual superiority. The poems also describe the annulment of these values, either by modernity or by physical leavetaking. The feminine rhymes, which, as Corcoran stresses, evoke the notion of loss, poignantly mirror such processes.

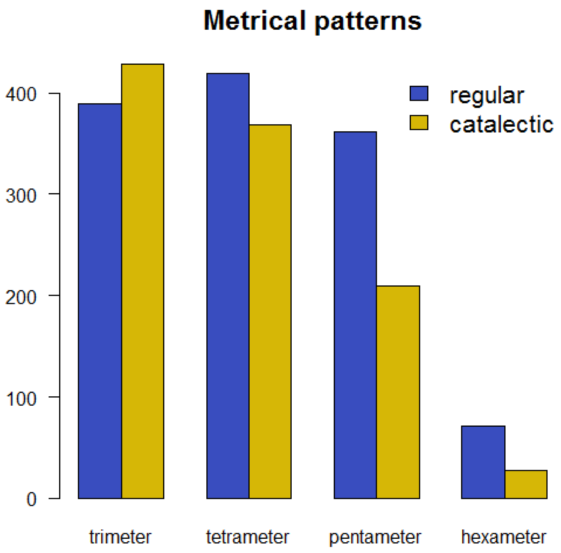


Figure 8.11. Absolute counts of regular and catalectic trimeters, tetrameters, pentameters and hexameters

As was noted above, having data about all the texts in a corpus enables scholars to investigate whether or not characteristics observed in individual poems are also representative of the full corpus. Traditional close reading tends to focus on a

[535] Terence Brown, *Louis MacNeice: Sceptical Vision*, p. 152.
[536] Ibid., p. 119.

limited number of texts, but by making use of computation, scholars can investigate the profusion of phenomena in a more circumferential manner. Digital methods may be used, for instance, to verify Robin Skelton's claim that many of MacNeice's metrical patterns are irregular. Skelton argues that the preference for an odd rather than an even number of syllables is a reflection of the poet's Celtic background. He uses the catalectic patterns in "The Sunlight on the Garden" to illustrate his argument. Allan Gilles notes, in a similar vein, that the poem "All Over Again" contains many skewed lines because of the addition of an extra stress.[537] To investigate the claim that MacNeice has a preference for irregular metrical feet, both the regular and the catalectic varieties were counted of all the verse lines whose metrical patterns could be classified automatically. Separate counts were also produced for the various line lengths. As can be seen, the catalectic pattern is used more frequently than the regular pattern in the case of trimeters only. In all other line lengths, the regular meters surpass the catalectic patterns. The diagram does clarify, however, that, although the catalectic lines do not dominate in absolute numbers, these incomplete patterns are still used highly insistently. To be able to interpret the results more thoroughly, nevertheless, data on the use of catalectic verse lines by other poets would clearly be valuable, as these could allow for a comparison of the use of incomplete metrical patterns among English and Irish authors.

In this case study, a number of experiments were also conducted to explore the relevance of sentiment analysis. Within the *Multi-Perspective Question Answering* (MPQA) tool, the *Harvard General Inquirer* package and the sentiment analysis tool that was developed by Bing Liu, lexicons have been developed with words that have positive and negative connotations. These lexicons have been made available separately, but, for the purpose of this study, these have been merged. On the basis of these combined lexicons, the tokens in this study's central corpus have been tagged. Subsequently, for each poem, the number of words with a positive and a negative connotation have been counted. The counts were normalised by dividing these by the total number of tokens. Next, the values for the negative words were deducted from the values for the positive words. Totals that were higher than zero were understood to indicate a chiefly buoyant register of speech, and negative values indicate texts that are overtly negative in tone. This relatively simplistic approach clearly has shortcomings. Words can express differing degrees of negativity, but, in this study, all positive and negative words were counted equally. Secondly, this analysis was based on counts individual words in a context-independent manner, and it failed to take into account the effect of negations. The simple counts of words from the lexicons can help, nevertheless, to produce a rough approximation of the connotations of the vocabulary.

[537] Alan Gillis, ""Any Dark Saying": Louis MacNeice in the Nineteen Fifties", p. 118.

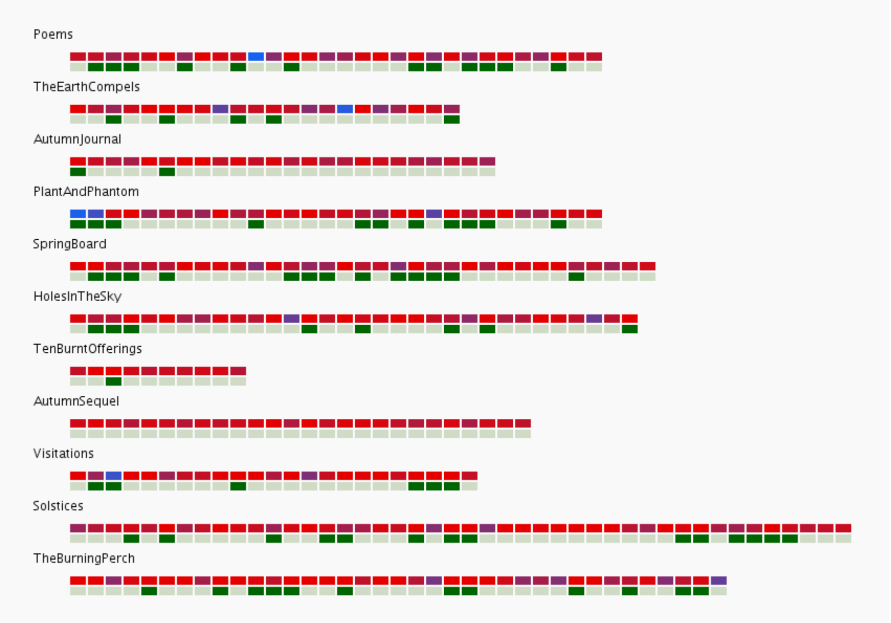


Figure 8.12. Results of sentiment analysis performed on all poems in the corpus

The results of the analysis are shown in figure 8.12. In this diagram, the blocks represent individual poems, and the overall sentiment of each poem is indicated by colours in a gradient scale. A predominantly negative vocabulary is shown using a red colour, and poems with a positive tone are rendered blue. The diagram also displays information on the themes of the poems, which had been assigned manually. If a poem was found to concentrate on decidedly negative themes, such as doom, nihilism and alienation, this was represented by a green block, shown directly below the bar that indicates the result of sentiment analysis. While, in most cases, the sombre themes are also described using words with negative connotations, as expected, there are also a number poems in which the sinister or morose themes are combined with mainly positive terms. Examples of poems in which a positive vocabulary coalesces with negative themes include “Wolves”, “Prognosis” and “Stylite”. The poem “Wolves” concentrates on the futility of critical reflection and evokes a fear of “the wolves of water”, which, as in many other poems, represents a nothingness and a nihilism. The text contains positive words such as “laughter” and “want” as it also proposes an antidote to nihilism. One solution is to unite with fellow human beings, and to imagine that dangers do not exist. The fear of disorder and of disintegration is also stressed effectively by the use of semi-rhyme. “Prognosis” is set during the end of a winter, and the persona reflects on what the new spring will bring. The poem contrasts a number of favourable and disagreeable options. Since the poem ends with a negative outcome ("Or will his name be Death / and his message easy"), the poem as a whole evokes a sinister sense of doom, connected to the imminence of the Second World War. The persona counteracts this fear by a belief in an alternative, more pleasurable reality. "Stylite" similarly presents opposites. It contrasts the central tenets of Christian religion, which demands austerity and a denial of earthly pleasures, with Greek philosophy which fully embraces the physical world. The positive words are mostly used in the description of the Greek god, which sits opposite the hermit on the stele. The poem as a whole describes a solipsism, and an isolation of the self from the world. A general pattern that may be observed is that poems with negative themes may also contain imaginative evocations of optimistic alternatives which counteract the negativity.

Figures 8.8 and 8.9 above list the poems with the highest normalised counts for specific forms of rhyme. Such analyses, which concentrate on literary devices in isolation, can often help to clarify the effect of these literary techniques within specific texts. As poems frequently contain salient combinations of different forms of rhyme, it can also be instructive to explore the values for these various forms of rhyme simultaneously. Figure 8.13 lists the texts which contain the high values for perfect rhyme, in combination with the counts for assonance rhyme and consonance rhyme.[538] The corpus contains 16 poems in which all the lines rhyme perfectly, and these have been ignored. [539] When a high percentage of one type of rhyme co-occurs with a relatively low percentage of another type of rhyme, this may warrant a more detailed close reading, as it is frequently the case that the lines that deviate in notable ways from a pattern which is otherwise regular are highlighted for a particular purpose. A clear example is "Autobiography", which has a score of 0.85 for perfect rhyme. All couplets contain perfect rhyme, with the only exception of the last couplet: "I got up; the chilly sun / Saw me walk alone". It may be argued that, while the stanzas that rhyme perfectly evoke the innocence and the simplicity of childhood, the breach of this pattern at the conclusion of the poem emblematises the psychological damage with which the poet entered adulthood.

Instances of perfect rhyme and slant rhyme have likewise been combined effectively in many of the other poems which have been identified in figure 8.13. In "The Grey Ones", the lines that rhyme perfectly evoke a traditional and increasingly outmoded form of living. Peter MacDonald argues that the state of greyness functions "as the drab background to modernity".[540] The evocation of modern life in

[538] The 20 poems with the highest values for perfect rhyme did not contain any semi-rhymes.

[539] The poems in which all the lines rhyme perfectly are "Chess", "Greyness is All", "Birthright", "The Revenant", "Beni Hasan", "Whit Monday", "The Ear", "Figure of Eight", "Mutations", "Passage Steamer", "Place of a Skull", "The Heated Minutes", "Bar-Room Matins", "Bluebells", "The Mixer" and "To Hedli".

[540] Peter McDonald, "Louis MacNeice: The Burning Perch", in: Neil Roberts (ed.), *Companion to Twentieth-Century Poetry*, Hoboken: Wiley-Blackwell 2003, p. 496.

the lines “Sprawled against the Gates of Doom / Whence all kebabs and cockstands come” (ll. 33-34) discernibly intrude into the surrounding lines. In the short poem “To Posterity”, the first line (“When books have all seized up like the books in graveyards”) is isolated from the rest of the poem because it is not part of the perfect rhyming scheme. This exclusion of the lines that describe traditional books stress the central fear in the poem that communication via written texts will be replaced by non-verbal media. In “Off the Peg”, line 8 (“Chameleons can adapt to whatever sunlight leaks”) is the only line which does not rhyme perfectly with any of the other lines. The poem as a whole describes the sensation that hackneyed phrases and worn-out habits can occasionally feel new and very relevant. The line which breaks with the regular pattern is also the point at which the poem begins to discuss the revitalisation of clichés.

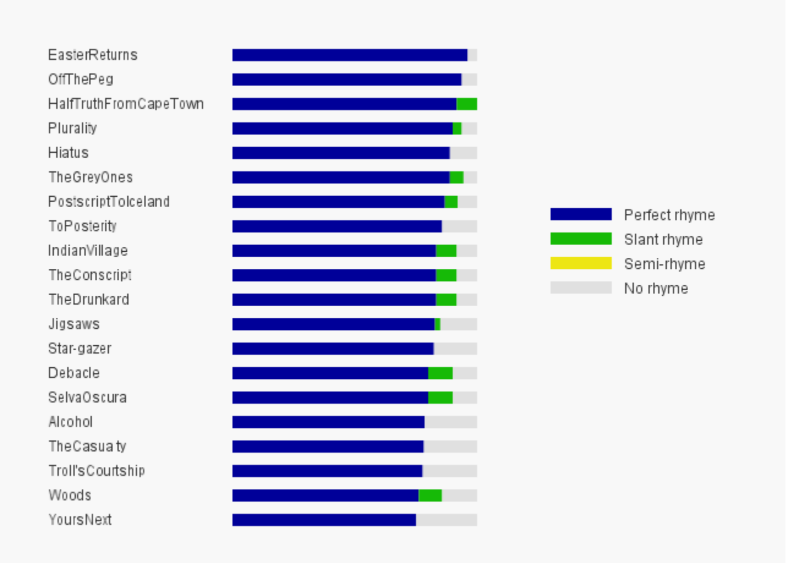


Figure 8.13. Percentages of lines that contain perfect rhyme, slant rhyme and semi-rhyme. The list was sorted using the values for perfect rhyme

The list which is shown in figure 8.13 is headed by “Easter Returns”. The poem focuses on the increasingly profane nature of religious festivities. The last word of each stanza rhymes with the first line of the stanza that follows. The last stanza forms the only exception to this regular structure. The schism in the formal structure of the poem suggests that, although traditional religious practices can no longer be followed, man can still find new forms of religious inspiration. Through such changes, man may experience that “the myth returns” (l. 22). The function of the rhymes and the non-rhymes in “Easter Returns” can be clarified further by linking this poem to “Place of a Skull” and to “Whit Monday”, which exclusively have lines which rhyme perfectly. These poems, like “Easter Returns”, focus on a loss of religion. “Place of a Skull” recounts the banal and irreverent reaction of the Roman soldiers to the death of Christ. The fact that the coat that was found on the dead body did not fit the soldiers epitomises the shift to a more nihilistic and a more secularised worldview. “Whit Monday” emphasises the idea that religious traditions and creeds have lost their relevance and that they fail to banish anxieties caused by modern life. These three poems on the disintegration of Christian faith effectively use regular perfect rhymes to underline the spiritual emptiness of religious practices.

A similar diagram can be produced to identify poems in which high values for slant rhyme have been combined with other forms of rhyme. One poem in which occurrences of slant rhyme alternate regularly with lines that contain perfect rhyme is “The Habits”. The poem has a very intricate structure, consisting of five stanzas, of five lines each. Two of these lines end in words that rhyme perfectly (“best” and “dressed”). Each stanza also has one line that ends in the word “habits”. The second, third and fourth stanzas contain two additional lines whose final words are loosely connected to the word “habits” via assonance rhyme (“carried”, “hypodermic”, “affidavit”, “sjambok”). The alternation between predicative perfect rhymes and the more uneasy assonance rhymes suggests that personal habits and social conventions can both be comfortable and vexatious. The regular pattern is discontinued in the final stanza. The absence of rhyming words in the poem’s last lines undergirds the notion that, in later life, these enforced habits “[o]utstayed their welcome” (l. 22).

Slant rhyme and perfect rhyme have also been combined intriguingly in “Dogs in the Park”, which centres around an opposition between the dog owner’s desire to control and to tame the dogs on the one hand and the savage and freedom-seeking dogs on the other. These stanzas have two lines which rhyme perfectly and two lines which contain slant rhyme or no rhyme. This structure reflects the tension between the urge to impose order on the one hand, and the desire to break away from this control on the other. Another poem which contains conspicuous combinations of slant rhyme and perfect rhyme is “The Rest House”. The poem, according to Terence Brown, evokes a scene which is “nightmarishly alive” and in which “objects have an unpleasant, spontaneous life of their own”.[541] The notion that the room has normal objects which nevertheless behave unexpectedly is bolstered by the form of the poem. “The Rest House” consists of two stanzas, and they both contain one instance of perfect rhyme (“veranda” and “Uganda” and “much” and “such”). The strangeness of the imaginative world is mimicked by the words which are coupled loosely via slant rhyme: “window” and “river” in the first stanza and “splintered”, “filtered” “children” and “extinguished” in the second stanza.

[541] Terence Brown, *Louis MacNeice: Sceptical Vision*, p. 141.

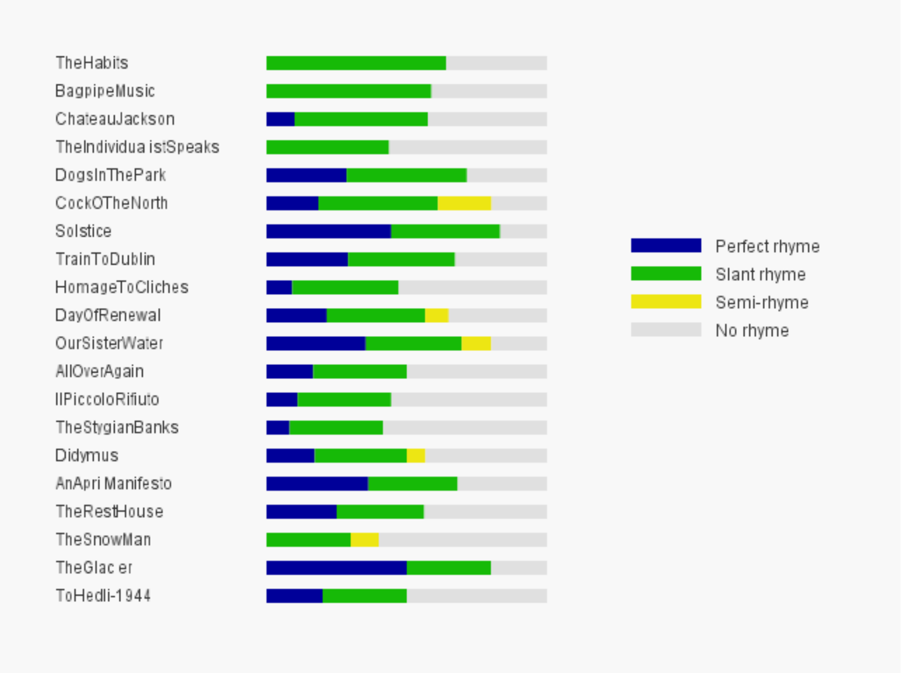


Figure 8.14. Percentages of lines that contain perfect rhyme, slant rhyme and semi-rhyme. The list was sorted using the values for slant rhyme

The various data values that have been produced about the formal aspects of MacNeice's poetry can be used productively to assemble texts which make use of the same literary techniques, and within such clusters of poems, scholars may investigate whether or not these poems also convey related ideas. Such considerations of the differences and similarities between texts can often quicken an interpretative reading of these texts. In the analyses that have been discussed above, poems have only been clustered on the basis of one or two variables, but, as this study has produced values about a wide range of different textual aspects, the similarities between different texts can also be investigated in a more encompassing manner. In this study, I have developed a method for the calculation of similarities between poems which was based on a concurrent analysis of 15 variables. For each poem, a vector was produced listing the values which were calculated for perfect rhyme, internal rhyme, alliteration, consonance, assonance, consonance rhyme, assonance rhyme, semi rhyme, deibhide rhyme, aicill rhyme, consonantal repetition, paronymy, iambic rhythms, trochaic rhythms and anaphora. The values that have been generated in this study could not all be compared directly, however, as a result of the fact that the different values have also been normalised in different ways. Values in the data set reflected either the number of counts per item, or the number of counts per verse line. To address these complications arising from these different units of measurement, all values have been rescaled through the calculation of z-scores.[542] An important advantage of working with z-scores is that they are essentially dimensionless. They are calculated by dividing all distances from the mean by the standard deviation. Both of these values are without a unit of measurement.

A first attempt at identifying related poems was based on a calculation of the cosine similarity for each possible pair of poems within the corpus, using the values for the 15 variables that were mentioned. Cosine similarity expresses the degree of similarity between two separate vectors as a value in between zero and one. Using this method, it was found that 118 pairs of poems displayed a cosine similarity of 0.95 or higher. On closer inspection of these results, it was found, however, that these cosine similarities were often high for poems which share a value of 0 for devices such as paronymy, semi-rhyme, and anaphora, which have only been used in a limited number of poems. In this study, the aim of identifying formal similarities was ultimately to explore whether or not the formal aspects of the poems correspond to what the poem intends to convey. Given this objective, it was more fruitful to concentrate on texts in which these literary techniques have actually been used, and to disregard poems without any occurrences. For this reason, an alternative method has been developed. The analytic algorithm that was used to produce figure 8.15 firstly calculates, for each pair of poems, the difference between all 15 values. Values were considered to be similar if the absolute difference was found to be less than 0.05, and, importantly, if the values for these variables were higher than 0. The network in figure 8.15 shows the poems in which at least three variables have such matching values. In these calculations, I have decided to leave the long compositions *Autumn Journal* and *Autumn Sequel* out of consideration.

Figure 8.15 connects poems which are formally similar. “Mayfly” and “Day of Renewal”, for instance, both contain internal rhyme: “One only day of May alive beneath the sun” (“Mayfly”, l. 4) and “Blood cholera flies blank eyes becoming forty” (“Day of Renewal”, l. 32). They have similar percentage of alliterating words, and they both contain aicil rhyme: “Flowers in the sun beside the jewelled water / Daughter of the South call the sunbeams home” (“Mayfly, ll. 6-7”) and “Claiming a different birthplace a wild nest / Further more truly west on a bare height” (“Day of Renewal”, ll. 13-14). There is also a thematic connection between the two poems. “Mayfly” muses on the transience of valuable experiences, using the brevity of the lifespan of mayflies as a poignant metaphor. The short lives of the mayflies, which are nevertheless stretched out “taut and thin and ringing clear”, inspire an urge to

[542] Z-scores indicate the distance from the mean of the data-set, expressed in standard deviations. As such, they offer a good indication of the position of a particular value within the entire distribution. See Roger Bilisoly, *Practical Text Mining with Perl*, p. 193.

savour precious moments by regarding these as moments which exist outside of time. At the same time, the poem admits that such permanence and timelessness are impossible, emphasising that such experiences in which time is made “elastic” can only exist as the result of a creative transformation. Like “Mayfly”, “Day of Renewal” reflects on the passing of time. The poem describes a number of past birthdays and notes that the human mind can transform events and locations into something permanent and static, while these continue to evolve in reality. A place which is recollected “in itself has changed but in our mind / Does not become but is” (ll. 27-28). “Day of Renewal” recommends a response to the transience of life which differs from the advice which is given in “Mayfly”. While the latter poem urges us to counteract the “pathetic fallacy of the passing hours” by using our imagination, “Day of Renewal” argues that evolution and moments of renewal need to be accepted, as these are essential to being alive. It stresses that “[d]eath is” and that “life becomes” (l. 19).

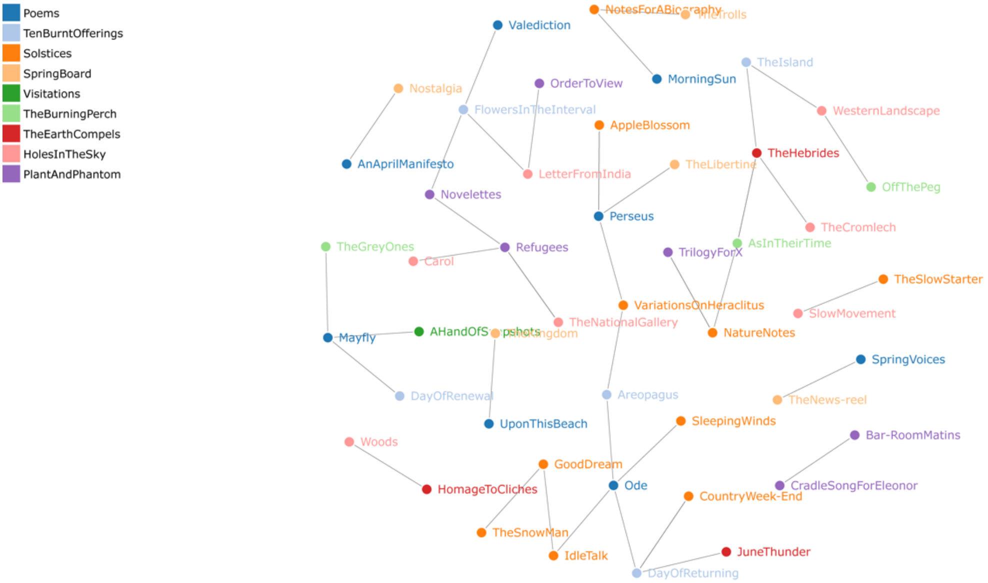


Figure 8.15. Similarities between poems, based on a concurrent analysis of 15 variables

Although systematic explorations of the formal similarities between texts cannot expose the thematic or the conceptual links between poems in themselves, such comparisons of linguistic properties or of literary techniques can often lead to juxtapositions of texts which may not have been related otherwise. Reading poems against the background of other poems can often further their interpretation. Figure 8.15 interestingly lists many poems which, next to their formal similarities, also share certain thematic concerns. “Cradle Song for Eleonor” and “Bar-room Matins”, for example, both concentrate on escapism. The speaker in the first poem urges the listener to look away from the “pity of it all” (l. 22). “Bar-room Matins” is written against the backdrop of the Second World-War and likewise describes a conscious but alcohol-induced effort to ignore the “doom tumbling from the sky” (l. 22). While the algorithm that was developed to establish similarities between poems did not consider their vocabulary, it coincidentally connected two poems whose titles share an adjective. “Slow Starter” and “Slow Movement” both contain trochaic lines, they have the same percentage of modal verbs and they have the same value for consonance. “Slow Starter” centrally describes a conflict between two conceptions of time. The first view posits that there is always time in abundance, and assumes that good things will come to those who are patient. This trite view is challenged by a second view which stresses that time is scarce, and that it is important take decisive actions. “Slow Movement”, from *Holes in the Sky*, likewise describes two different ways of experiencing time. It depicts a persona who sits opposite an unknown woman who is asleep. The persona appears to image that he features in her dreams. The dream sequence appears to exist outside of time, with “yesterday cancelled out” (l. 3). The end of the dream, which is described in the final lines of the poem, also marks a return to the fast-paced time of external reality, “[a]ccelerando con forza” (l. 23).

Striking commonalities were also found between “Homages to Clichés” and “Woods”. “Homages to Clichés” initially presents the social conventions that make interpersonal communication predictable as comfortable and appealing, but, as the poem progresses, the clichéd interactions increasingly begin to amplify the distances between the two persons who converse. The stale and banal conventions which are enacted also represent an inability to reflect critically and originally, and the poem eventually expresses a longing for a radically nonconformist world in which “[n]ext year is this year, sometime is next time, never is sometime” (l. 79). “Woods” contains a very similar progression from glorification to debasement. While the poem originally presents the English woods as kingdoms “free from time and sky” (l. 10), the penultimate stanza also accentuates the insipid and artificial nature of the tidy woods which were “assured / Of their place by men” (l. 25). “Homages to Clichés” and “Woods” both focus on controlled and predictable phenomena which can be satisfying to some extent, but which ultimately fail to stir meaningful or profound emotions.

Analyses of the similarities between poems can help to expose conspicuous aspects of MacNeice’s poetic language. The methods that were developed for this study did not primarily concentrate on the semantic contents of MacNeice’s poetry, but on devices that are based on sounds. Edna Longley stresses, nevertheless, that MacNeice had a "continual drive to reconcile the demands of form and content".[543] Perhaps to an even greater extent than some of his contemporaries, MacNeice was deeply interested in the complex relations between sound and meaning. As has been shown in this chapter, a systematic analysis of the sonic patterns in the various verse lines can productively fortify an interpretative reading of MacNeice's poetry.

8.3. Conclusion

8.3.1. Discussion of the case study

The case study that was conducted for this study primarily aimed to contribute to a general understanding of the concrete challenges that can emerge during processes of data creation and data analysis. Next to this main objective, however, it has been shown additionally that the various methods that are associated with literary informatics can effectively be applied to generate relevant new insights about the poetry of Louis MacNeice. In "Postscript to Iceland" MacNeice writes that his urge to find those texts that can truly inspire is frustrated by the feeling of being inundated by the many works that are available:

Rows of books around me stand,
Fence me round on either hand;
Through that forest of dead words
I would hunt the living birds

The stanza may simultaneously be read as an admission of the difficulty of finding a unique poetic voice amidst the literature produced by previous generations. This study aimed to chase "the living birds" in MacNeice's own literary oeuvre by making use of algorithms. Computers can detect patterns within texts fragments with an acute exactitude and with an implacable consistency, and, utilising this capacity, this case study was able to disclose a number of aspects of MacNeice's poetic language which had not been discussed yet in earlier studies. I have shown, for instance, that MacNeice often used alliteration in poems with an escapist surge, or in poems which ruminate on the distance between the self and the external world. The computer's ability to detect patterns also helped the disclose the fact that MacNeice frequently made use of specific combinations of alliterative words, in which two pairs of sounds are repeated, or in which one pair of sounds is nested within a second pair of sounds. The methods that were discussed in this thesis also facilitated a systematic examination of all occurrences of semi-rhyme and deibhide

[543] Edna Longley, "Louis MacNeice: Aspects of His Aesthetic Theory and Practice", p. 151.

rhyme. The type of sonic agreements these devices are based on are often difficult to notice for human scholars. These forms of rhyme frequently occur in poems which concentrate on stark oppositions, or on the petrification or the inertia that results from an unruly plurality. Deibhide rhyme is often used, more specifically, to describe social distances. This study has also demonstrated that the poems which lack consonance all describe a solitude, and that poems about the loss of religion often make a heavy use of perfect rhyme. None of the critical studies that have been investigated for this study, moreover, have commented on the importance of money, war and food imagery.

Computational methods can be applied effectively to investigate whether or not the qualities which earlier scholars have observed within singular text are also present in other texts. In this study, such critical expansions have often resulted in new observations. Robin Skelton has discussed the nature and the function of Celtic poetic devices such as deibhide rhyme and aicill rhyme, illustrating his explanation via a close reading of the two poems “Aubade” and “Order to View”. Using computation, it was possible to study the profusion of these specific literary devices within MacNeice’s entire literary career. In combination with a more comprehensive assessment of the poet’s use of catalectic metrical patterns, such methodical examinations or deibhide rhyme and aicill rhyme can lead to a sharper delineation of the Celtic influences in MacNeice’s work, and, ultimately, to an improved understanding of the poet’s relation to Ireland and to the Anglo-Irish tradition. Along similar lines, Neil Corcoran’s observation that the feminine line endings in “Leaving Barra” evoke a melancholy sense of loss formed the incentive for a wider exploration of all the poems in which the majority of lines have feminine endings. Interestingly, it was found that the poems whose line endings are almost exclusively feminine all depict islands. Computation can also be used to supply new forms of support for observations made by previous critics. Critics generally agree, for instance, that *Autumn Sequel* and *Ten Burnt Offerings* are distinctive within MacNeice’s oeuvre because of its concentration on philosophical topics, and because of the absence of engaging imagery. The case study has indicated, furthermore, that the poems from the early 1950s have a lower type-token ratio, and that they generally contain words with a lower number of syllables.

The case study that was conducted for this thesis did not aim to falsify the results of earlier studies. As was explained, most of the algorithms were inspired by observations which had been made in earlier studies, and the aim was often to extrapolate from claims which were made about individual texts. As the emphasis was primarily on occurrences and frequencies of literary devices, this study did not produce interpretations which were radically different from previous critical studies of MacNeice’s work. Algorithmic criticism, of the form that was illustrated by the case study, primarily aims to invigorate interpretative readings which, in line with the traditional close reading method, are based directly on the formal characteristics of the text. Computational methods can effectively identify texts with distinctive formal or structural properties, and such information about their uncommon or their unique properties can often be used productively in critical assessments of texts. Digital methods cannot of themselves add interpretation, and for his reason, this study did not result in strong refutations of earlier interpretations. The toolset that is associated with algorithmic criticism cannot independently establish, for instance, that a poem which has conventionally been viewed as good literature is in fact of a poor literary quality. Statistical methods can occasionally lead to reassessments of texts, nevertheless. Many of the analyses that were conducted for the case study highlighted the statistical particularity of the poems ”Place of a Skull", "Memoranda to Horace" and "Spring Sunshine". These poems have not been discussed at all in Edna Longley's extensive monograph on MacNeice's verse. Evidently, an unusually high or an exceptionally low frequency of literary devices does not necessarily imply that these devices have also been used in an interesting or in befitting manner. As qualitative methods allow for a more embracive and a more dispassionate form of research, they can prompt scholars to ask the question whether or not particular poems have been neglected deservedly.

8.3.2. Discussion of data analysis

In Chapter 4, it was argued that literary informatics mainly buttresses three basic scholarly primitives: annotation, comparison and discussion. Chapter 6 has discussed the main difficulties which can arise during the creation of annotations. This chapter has focused in more detail on the many different ways in which digital methods may support comparison and discovery. Using the practical work that was undertaken for this study as inspiration, two different types of comparison can be distinguished. Analytic methods can firstly compare the stylistic qualities of groups of texts in their entirety. In the case study, I have applied various analytic methods to explore the differences and similarities between 11 volumes of poetry, focusing more specifically on the differences between the volumes which were written during the early 1950s, and the volumes which were before and after this period. Comparative analyses may focus, secondly, on the differences and similarities between singular texts. When the results of analytic methods contain references to individual poem titles, these poems can subsequently be analysed more closely, using the information about the distinctive qualities of these texts as reading guides. In Chapter 2, literary texts have been described as linguistic compositions in which particular recondite concepts have been translated into formal features. If it is shown that a poem makes an extensive use of a specific literary device, such an observation may often be explained by relating some of the connotations of this literary device to the central thematic concerns of the text which contains this device. The statistical artefacts which are constructed, importantly, do not contain interpretation in themselves. The resources that result from algorithmic processing function as means rather than as ends, as these mainly form the materials with which scholars can fortify their interpretation.

Scholars who aim to compare texts can commonly choose from a broad range of statistical techniques, and it can often be taxing to select an appropriate method. The differences between two distinct classes of texts may be examined using supervised learning techniques, of which Student's t-test, logistic regression and Naieve Bayes all form concrete examples. When scholars aim to subdivide a corpus into smaller clusters, they can make use of k-means clustering, calculations of Euclidean distances, PCA or nearest neighbour analyses. These different methods are all based on different algorithms, and they consequently produce different results. Such differences can be subtle in some cases, but also quite dramatic in other cases. Even when scholars have decided to make use of one particular technique, they frequently have the possibility to influence the results by varying some of their initial parameters. In the case of classification, the results of the analyses can often be manipulated directly by varying the sizes of the training sets and the test sets. In this study, it was found, for example, that the nature of the network diagram displaying formal similarities between poems can change dramatically along with the variables which are considered.

Within the context of algorithmic criticism, there are few guidelines which can help scholars to make a reasoned selection. Statisticians can generally characterise the varying level of accuracy that can be achieved by these analytic algorithms, by making use of measures such as the precision and recall ratio, and the residual sum of squares. In studies which aim to assign an author to an unassigned text, or in studies which aim to date a particular text, the accuracy of a method can be verified through tests based on texts whose authors are known. This concept of validity seems appropriate in studies which operate under the scientific assumption that questions can be answered conclusively through a single answer. In the context of algorithmic criticism, however, such strict criteria for the evaluation of the adequacy of analytic methods can seldom be applied. In this specific discipline, the results of statistical procedures can be considered relevant when they ignite new ideas about the text that are studied, or when they convincingly stimulate hermeneutic interactions. A method which performs poorly according to statistical accuracy measures can still be valuable for literary criticism if it helps scholars to discover striking formal or thematic similarities between two texts. Such new ideas can be fostered when statistical methods manage to expose striking differences between different sets of texts, or when statistical methods can dispose formal similarities between texts which, according to the scholar, also share particular thematic concerns. During the case study, I experienced that many of the methods that were applied failed to inspire such new ideas. The statistical techniques often have to be used in an aleatory or experimental fashion, and it is often difficult to know beforehand whether these analyses can yield interpretable results. The choice of a particular method often reflects an individual scholar's preference.

Next to the fact that this multitude of statistical methods can complicate an effective comparison of texts, the discovery of relevant texts can likewise be challenging because of the absence of a single reliable heuristic method. In the case study, I tried to find interesting examples of the use of semi-rhyme or deibhide rhyme by sorting all the poems according to the frequency counts of these devices. The poems in which devices have been used most profusely do not necessarily contain the most interesting instances. The observation that a particular fragment has literary significance demands a qualitative assessment, and it continues to be challenging to detect such significant cases solely on the basis of quantitative methods. In *Modern Poetry*, Louis MacNeice wrote that “the rules or ‘laws’ of poetry are only tentative devices” and that “[t]here is no Sinaitic recipe for poetry, for the individual poem is the norm”.[544] MacNeice’s explanation implies that, because of the intrinsic uniqueness of each poem, it is exceedingly difficult to develop steadfast rules for the identification of the qualities that can make a poem extraordinary. Because of its fortuitous and unpredictable nature, the discovery of all relevant texts is difficult, in traditional and in computer-based research alike. It can be expected, nevertheless, that the methods of algorithmic criticism, which can recognise occurrences of many literary devices, and which, despite specific challenges, can systematically compare the texts in a given collection, can ultimately help scholars to make such serendipitous discoveries more quickly.

In the scholarly writing about the marriage between literary studies and informatics, one crucial and fundamental question reoccurs insistently: do computational methods essentially enable scholars to perform the same set of tasks, albeit with more speed or with more precision, or can these various technologies genuinely effectuate fundamental transformations concerning the knowledge that can be produced? The case study that was conducted for this thesis has shown that digital techniques, because of their unparalleled capacity to methodically detect all occurrences or literary devices, can radically expedite and ameliorate parts of the traditional methodology of literary criticism. They can vastly magnify the range of textual aspects that can be observed within literary works. It is not immediately clear, nevertheless, that such heightened forms of perception can veritably qualify as a fundamental epistemological change. Martin Mueller has argued convincingly that epistemological innovation is often a gradual and incremental process, and that many piecemeal changes, occurring resolutely over the course of a longer period, can ultimately amount collectively to a potent “change in kind”.[545] This study has assumed analogously that, although the methodological renewal that can be instigated by computation does not instantaneously result in a profoundly different form of knowledge, many cumulative methodological emendations can eventually result in a palpable transformation of the nature of literary research. Whereas the results of computer-based research rarely produce draconian disciplinary transformations directly, the capacity of digital methods to recognise patterns in the usage of literary devices can be of clear benefit to the field of literary

[544] Louis MacNeice, *Modern Poetry: A Personal Essay*, p. 33.

[545] Martin Mueller, “Digital Shakespeare, or towards a Literary Informatics”, p. 286.

criticism. Scholars with a genuine interest in literature ought to eagerly embrace every additional instrument which can shed new light on sinuous and multifaceted literary texts.

Chapter 9

Visualisation

9.1. Introduction

One of the most striking consequences of the advent of digital methodologies is the marked upsurge in the use of visualisations, and, simultaneously, an attenuation of the centrality of text. Conventional humanities research is dominated profoundly by the written word. Especially in fields such as literary studies and philology, text is frequently both the object and the outcome of academic research. Unlike research projects in the natural sciences, which tend to rely heavily on quantitative data, works of literary criticism have rarely employed visual displays such as graphs and charts for the organisation and the dissemination of knowledge. The growing use of visualisation is spurred both by the nature of the digital medium, and by the type of research that the digital medium makes possible. Firstly, while it is obviously possible to transfer images via paper-based books or journals, the ease with which images can be created, edited and disseminated in a digital environment is unequalled. In printed publications, texts and images are the only modalities which can be conveyed, but this limitation no longer exists in the digital space. The computer is in fact "a continuum in which many signs of representation can happily co-exist".[546] On the computer, it is often unclear "where the pictorial space ends and the verbal space begins".[547] Secondly, as was discussed in earlier chapters, the essential plasticity and computability of digital text has also inspired innovative forms of analysis, and the growing interest among literary scholars in quantification and in statistics has also urged scholars to explore whether or not the salient features of numerical data sets can be clarified and communicated effectively though images.

In most cases, the outcomes of quantitative analyses can be examined most effectively if these are presented visually. Statistical manipulations of quantitative data initially result in new numbers, but bare numbers displayed in a tabular form often fail to facilitate an efficient exploration of trends or of notable exceptions to these trends. An adroit use of visualisation techniques can help scholars to promptly discern relevant characteristics of voluminous data collections. The

[546] Jay Bolter, *Writing Space: The Computer, Hypertext, and the History of Writing* (Hillsdale N.J.: L. Erlbaum Associates 1991), p. 60.

[547] Ibid., p. 74.

transfer of ideas via graphical displays derives much of its attraction from the fact that it is often perceived as a more immediate form of communication. Dehaene explains that viewing objects requires considerably less neural activity than reading.[548] Unlike the capacity for vision, the human brain does not have an innate wiring for processing written texts, and the ability to read demands numerous new connections between distinct areas in the brain. Stan Ruecker stresses similarly that human beings have a great capacity for perceiving and processing visual information, as the human brain has "ecological advantages" that enable it to process "environmental features such as luminosity, motion and colour" with great speed.[549] Merely from glancing at a diagram or a graph, human viewers can identify notable features almost immediately, and, in many cases, the focus is directed intuitively to outliers, colour contrasts, or other irregularities in the overall shape, if present. This visual orientation and the innate aptitudes for perceiving patterns and regularities can be leveraged expediently in the creation and the communication of knowledge.

Since studies in the field of literary criticism generally aim to illuminate the meaning of texts, Sinclair et al. argue that the value of scholarly tools for the visualisation of literary texts can be gauged "by determining how well it supports this interpretative activity".[550] This chapter concentrates more closely on the ways in which visualisation techniques can contribute to hermeneutic processes. This chapter initially offers a definition of data visualisation. Next, a classification is proposed of the various ways in which data about texts can be clarified visually. Finally, an assessment is made of the ways in which the patterns that emerge within visualisation can support or obstruct interpretation.

9.2. Definition

Providing a clear definition of the term 'visualisation' is complicated by the fact the word may refer to a wide range of visual phenomena. Following Manovich, a distinction can be made between visualisations which represent objects or phenomena directly and visualisations which primarily represent data about these objects or phenomena. In the first case, graphic displays are essentially renditions of the "a priori fixed spatial layout of the real physical objects such as a brain, a coastline, a galaxy".[551] These types of visualisations, which Manovich refers to as 'scientific

[548] Stanislas Dehaene, *Reading in the Brain: The Science and Evolution of a Human Invention* (New York: Viking 2009).

[549] Stan Ruecker, "Rich Prospect Browsing Interfaces", in: *Affordances of Prospect for Academic Users of Interpretively Tagged Text Collections*, University of Alberta 2003, p. 5.

[550] Stéfan Sinclair, Stan Ruecker & Milena Radzikowska, "Information Visualization for Humanities Scholars", in: *Literary Studies in the Digital Age*, Modern Language Association of America 2013, n. pag.

[551] Lev Manovich, "What Is Visualization?", in: *Visual Studies*, (2011), n.pag.

visualisations', are based on a graphical structure which has been observed in a particular physical environment. The existing shapes, colours and dimensions are reproduced in an abstracted form on a certain medium, often for the purpose of further analysis. Visualisations that mimic physical objects or phenomena can be distinguished from 'data visualisations' or 'information visualisations',[552] which generally involve "a mapping between discrete data and a visual representation".[553] Maureen Stone views data visualisation similarly as "the field of study that uses interactive graphical tools to explore and present digitally represented data, that may be stimulated, measured, or archived".[554] Information visualisations take a shapeless collection of data as input, and attempt to illuminate characteristics of this data set through a specific organisation or arrangement of graphic elements.

Using the adjective "visual" may be misleading, as written texts, like graphs and diagrams also consist of signs which are to be perceived through vision. The term 'data visualisation' can be defined more precisely as an operation through which specific properties of a data set are represented using the graphical modality. 'Modality', in short, refers to the manner in which the information that is conveyed through a medium is encoded. It consists of a distinct class of signs which senders and recipients can use to communicate. Modalities may be classified using the human senses that are needed to decode the information. Written text and images are both visual modalities, while music and braille are aural and tactile modalities, respectively. An important difference between texts and images is that the former uses arbitrary signs that form a linguistic code that must be learned, while images are available to anyone without formal instruction. A thorough discussion of the graphic modality is offered in Jacques Bertin's influential monograph *Semiology of Graphics*, which describes seven graphic variables, namely shale, scale, tonal value, texture, colour, orientation and location.[555] The graphic modality communicates knowledge through dexterous combinations of these basic variables. Leland Wilkinson, in *The Grammar of Graphics*, argues in a similar fashion that complex graphics can be broken down into a limited number of basic graphic primitives. He explains that graphics can be created by mapping a data set to a perceivable physical representation, and that such representations consist of "aesthetic attributes". Building on the categorisation that was developed by Bertin, Wilkinson enumerates a large number of aesthetic attributes, including include position, size,

[552] The two terms will be treated synonymously.

[553] Lev Manovich, "What Is Visualization?", p. 2.

[554] Maureen Stone, "Information Visualization: Challenge for the Humanities", in: *Working Together or Apart : Promoting the next Generation of Digital Scholarship : Report of a Workshop Cosponsored by the Council on Library and Information Resources and the National Endowment for the Humanities*, Washington D.C.: 2009, p. 44.

[555] Jacques Bertin, *Semiology of Graphics* (Madison: University of Wisconsin Press 1983).

shape, rotation, colour, saturation, orientation and blur. Wilkinson's attributes are grouped into five categories: "form, surface, motion, sound, and text".[556]

Data visualisation often involves a change in modality. The original objects which are studied generally employ a set of semiotic signs that differ from the signs that occur in the source that is ultimately represented by the visualisation. In the context of literary informatics, visualisation involves a process in which a message in the textual modality is represented and clarified through information encoded in the graphical modality. In previous chapters, it was explained that computer-based text analysis involves the application of a range of techniques which can convert a corpus of linear and discursive texts into discrete and structured data, and which can further analyse the resultant data set in a variety of ways. Visualisation is a specific type of processing in which the results of such analyses are subsequently represented via graphical primitives. Patterns created by variations in the use of the colours, shapes and locations are taken to represent differences and resemblances among the original texts.

In general, however, data visualisations do not consist exclusively of signs from the graphical modality. In his essay "The Rhetoric of the Image", Roland Barthes examines the various ways in which images can convey a meaning, and identifies three classes of messages.[557] The literal message, first, consists of the concrete objects or events which are depicted. Next to the literal denotation, the image also has a coded or symbolic message, consisting of the connotations of the objects that are shown. Importantly, Barthes stresses that images often contain a linguistic message, which is made up of the words which occur in the image itself, or in the caption or in the heading that accompanies the image. Barthes' analysis was based principally on the messages produced by advertisement photographs, and it cannot be extended directly to explain the rhetoric of data visualisations. One important difference is that the non-textual elements in a data visualisation cannot establish denoted or symbolic messages independently. Unlike scientific visualisations, graphical renditions of data do not directly portray concrete objects or phenomena. Barthes maintains that, since the signifier and the signified of the objects that are visible in photographs are "quasi-tautological",[558] the literal message can be referred to as "a message without a code". It may be argued that the graphical modality used in a data visualisation consists exclusively of codes which cannot produce a message in themselves. Specific permutations of graphical primitives principally convey patterns, exposing differences or similarities between various data values. The more precise meanings attributed to specific components of the graph are not codified and they are mostly unique to a given visualisation. They remain fickle and indeterminate until their particular signification is clarified via

[556] Leland Wilkinson, *The Grammar of Graphics* (New York: Springer 2005), p. 118.

[557] Roland Barthes, *Image, Music, Text* (New York: Hill and Wang 1977).

[558] Ibid., p. 154.

the textual modality. Since data visualisations are rarely self-explanatory, developers of visual displays generally need to add a legend or another form of textual support to explain what the various components of the graphic represent. The graphical modality thus depends on the textual modality to produce a meaning beyond the pattern itself. In the case of data visualisations, the text's ability to provide anchorage is vital, as, without it, graphical displays remain problematically devoid of meaning.

The nature and the function of visualisation can be clarified further by contrasting it with the function of typography. Whereas the application of typography and the use of visualisation both result in a certain visual presentation, there are also a number of important differences. The purpose of typography is to show the lexical codes of a text in a specific form, to clarify the logical structure of the text, and, more generally, to ensure that the text is legible and accessible. Bolter argues that typography ought to “make the letter unobtrusive”[559] and to let readers focus on the contents of the message in the textual modality with as little distractions or obtrusions as possible. It is misleading, nonetheless, to suggest that typography is a neutral layer. Typographers present the text in a particular manner which, to a higher or lesser degree, influences the way in which texts are experienced. Typography, importantly, does not distort the linearity of the text. Linguistic signs derive part of their meaning from their placement within a specific context, and, if the linear order of the tokens is forfeited, this mostly means that readers can no longer decode the message. Whereas typography is concerned with the visual appearance of the text in its full linear form, data visualisation focuses on the production of a succinct and a non-linear rendition of data about texts. The transition to the graphical modality demands operations such as tokenisation and quantification, and, in this process, the linearity of the text is crucially discarded. In an information visualisation, the text can no longer be read in its entirety and the focus is singularly on data about the text. While typography is generally meant to guide the reader’s attention effectively to the text and to its various logical components, data visualisation makes the text itself invisible and focuses exclusively on the forms that are produced out of data.

Stephen Few notes that graphical displays of data can be used for two purposes: “sense-making (also called data analysis) and communication”.[560] Visualisations may initially help scholars to explore the characteristics of large data sets. Once conspicuous aspects have been identified scholars may also generate graphs and diagrams to communicate specific ideas to peers. Graphic displays can either be a research tool or a means of communication. The next section argues

[559] Jay Bolter, *Writing Space: The Computer, Hypertext, and the History of Writing*, p. 63.

[560] Stephen Few, “Data Visualization for Human Perception”, in: Mads Soegaard & Rikke Friis Dam (eds.), *The Encyclopedia of Human-Computer Interaction*, Aarhus: The Interaction Design Foundation 2014, n.pag.

that, while visual displays can be highly effective during the exploratory phase of research, they also have a number of characteristics that complicate the communication of scholarly knowledge.

9.3. Expressiveness and rhetoric of visualisations

While scholarly knowledge is conventionally disseminated via articles and monographs, the growing use of visual materials increasingly undermines the position of text as the privileged channel for the transfer of facts and of ideas. In a discussion of the distinction between the terms "illustration" and "visualisation", Jessop asserts that "an illustration is intended merely to support a rhetorical device (usually textual)" and that a visualisation is critically "intended either to be the primary rhetorical device or serve as an alternative but parallel (rather than subordinate) rhetorical device". [561] Jessop's argument implies that data visualisations may partly or wholly supplant a textual publication. Visualisations are generally produced, like written texts, to transmit information, but an exclusive use of the graphical modality also complicates or obviates the transfer of particular types of messages. To understand the value of visualisation within the context of literary research, it is useful to concentrate initially on the means by which graphics can produce meaning, and on the type of messages they can encode.

An analysis of the components of visualisations may be based on conceptualisations of the visualisation process. Visualisations are created in various stages, and many of these stages can affect the manner in which graphics can convey information. Useful conceptualisations of the visualisation process are provided by Ben Fry,[562] by Leland Wilkinson and by Chen and Floridi.[563] In a similar fashion, Hullmann and Diakopolous distinguish four "editorial layers" that can impact the meaning of graphics.[564] Using these various descriptions of the visualisation process as inspiration, it may be argued that the meaning of visualisations is determined most fundamentally by four aspects. All descriptions of the visualisation process concur that graphics are based on data. The data roughly

[561] M. Jessop, "Digital Visualization as a Scholarly Activity", in: *Literary and Linguistic Computing*, 23:3 (2008), p. 283.

[562] Fry's textbook *Visualizing Data* presents a sequence consisting of seven stages, in which creators of visualisations sequentially acquire, parse, filter, mine, represent and refine data. Eventually, viewers may also be enabled to interact with the visual display. See Ben Fry, *Visualizing Data* (Cambridge: O'Reilly Media Inc. 2008), p. 8.

[563] Chen and Floridi divide the "visualisation pipeline" in steps such as "enriching & filtering", "visual mapping", "rendering" and "displaying". See Min Chen & Luciano Floridi, "An Analysis of Information Visualisation", in: *Synthese*, 190:16 (26 September 2012), p. 3422.

[564] The nature of a visualisation is determined by the data that are shown, the way in which these data are "mapped to the visual domain", the presence of annotations, and the degree of interactivity. See Jessica Hullman & Nicholas Diakopoulos, "Visualization Rhetoric: Framing Effects in Narrative Visualization", in: *IEEE Transactions on Visualization and Computer Graphics*, (2011), p. 4.

indicate the domain which is described by the chart or the graph. Data visualisations can only represent those resources which were classified in Chapter 4 as structured annotations or as derived data. Captured data are essentially digital reproductions of texts, and these cannot be mapped directly to the visual domain.[565] The data are visualised can represent data either about individual texts or about corpora in their entirety.

A second aspect which contributes to the meaning of the visualisation is the type of additional processing these data have undergone. The values may, for instance, be sorted, clustered or filtered. As was indicated, visualisations consist of a mapping between qualitative or quantitative data and a set of graphical primitives. Visualisations can be classified, thirdly, by considering how the values are mapped to the visual domain. The data may be represented using different aesthetic attributes,[566] including shape, rotation, colour or position. Selecting a particular graphical primitive may imply a loss of information, as the granularity with which such primitives may be varied does not necessarily match the granularity of the original data. The meaning of the various visual components may be clarified using legends or labels.[567] This third criterion is defined broadly, and also covers the usage of different types of scales and coordinate systems.[568] A fourth aspect which

[565] In his article "What is Information Visualisation?", Manovich argues that his original definition of information visualisation, as a mapping between discrete data and a graphical display, is complicated by the advent of a number of recent visualisation tools. Although Manovich does not use the term 'modality', he does stress that the reduction of information to an arrangement of graphical primitives is essential to the creation of a visualisation. On the basis of this argumentation, he concludes that a word cloud, for instance, is not a data visualisation. Manovich claims that, since the words that make up the text are also used directly in the visualisation, there is no actual reduction. A word cloud, consequently, is referred to as a 'direct visualisation'. It is a representation in which new forms are produced from the original media without any reduction and without a change in modality. This view, however, is abstruse, as it seems clear that word clouds and the captured data they are based on engage different modalities. While word clouds and natural language texts contain the same codes, these symbols evidently have different functions. The linear words order, which is necessary to understand the meaning of the original text, is, in most cases, discarded fully in the case of a word cloud. The symbols that are used in the latter type of display are primarily labels for metrics. In word clouds, differences in graphical variables such as scale and position are applied directly to the labels to clarify the values that have been calculated for these metrics. Manovich argues that "the two-stage process of first counting, or quantifying data, and then representing the results graphically" is no longer applicable in the case of direct visualisations. To create word clouds, however, it is still necessary to tokenise the text and to count frequencies of types. A word cloud is based on derived data and not on captured data.

[566] Bertin's term "graphical primitives" and Wilkinson's term "aesthetic attributes" are considered synonymous.

[567] The textual clarifications which are employed in a data visualisations can connected conceptually to what Barthes refers to as the "linguistic message" of the image.

[568] Wilkinson explains that the scale of the graphic indicates the ratio between the actual distances in the data collection and the new distances in the graphic. Visualisations may also be based on different coordinate systems, which consist of "sets that locate points in space". The Cartesian coordinate system is used most commonly.

may result in further modifications of the meaning of the visualisation is the degree of interactivity. Sinclair et al. explain that static visualisations are "fundamentally tools for display", and that interactive visualisations are also research tools, as they enable scholars to experiment with the different renditions of data sets. This exploratory process is often sequential and iterative, as researchers may "revisit previous steps at a later stage and make different choices, informed by the outcomes produced in the interim".[569] Whereas Jessop opines that interactivity ought to be viewed as a defining characteristic of visualisation,[570] this thesis will use the term to refer both to fixed representations of data and to visualisations which can be manipulated flexibly.

In recent years, a large number of tools have become available for the analysis of texts and for the visualisation of the results of such analyses.[571] The current profusion of tools potentially complicates the task of matching visualisation methods to specific research questions. Using the four aspects which were discussed, namely, the type of data, the processing these data have undergone, the graphical primitives and the degree of interactivity, the main functionalities offered by available visualisation tools can be described effectively. Tools such as Voyant Type Frequency Chart and the Distribution tool in TaporWare, for instance, focus on word frequencies via bar chart or via line charts. In both tools, only one file may be uploaded simultaneously, and this makes the tool more suitable for document analysis than for corpus analysis. The tools divides the text into segments, and presents the frequencies of terms within these separate segments. In both applications, the sizes of the text segments may be specified interactively.

In relation to the third aspect that was discussed, the visualisation's graphical primitives, a distinction can be made between diagrams which are based on conventional statistical models on the one hand and visual displays which explore new forms on the other. The first class of diagrams are based on visual models such as the bar chart, the line chart and the pie chart, which were first developed by William Playfair in the late eighteenth century.[572] These displays can normally be produced within statistical packages such as SPSS or R. A second class of visualisations clarifies data via innovative amalgamations of visual primitives. One example is Voyant Bubblelines, which is a tool for the creation of distribution graphs. It can clarify the frequencies of terms within distinct segments of individual texts. Voyant Bubblelines, interestingly, represents such values via the radiuses of circles.[573] Stefanie Posavec's project *Literary Organisms*, which visuali-

[569] Stéfan Sinclair, Stan Ruecker & Milena Radzikowska, "Information Visualization for Humanities Scholars", pp. 1–2.

[570] M. Jessop, "Digital Visualization as a Scholarly Activity", p. 283.

[571] The TAPOR directory of digital research tools lists more than 150 applications for the visualisation of data.

[572] Edward Tufte, *The Visual Display of Quantitative Information* (Cheshire: Graphics Press 1983).

[573] <http://voyant-tools.org/tool/Bubblelines/> (11 March 2014)

ses Jack Kerouac's novel *On the Road*, also provides a striking example of an innovative type of visualisation.[574] Each chapter of the book is represented as a flower, in which the size of the petals indicate the number of words. In addition, data were created manually about the various themes that are covered in each chapter, and these data about themes determine the colour of each flower.

While the informative value of standardised and innovative graphs may be roughly identical, their aesthetics often differ sharply. Hullmann and Diakopolous note that specific graphical primitives have both a denotation and a connotation. The former term refers to the manner in which graphical primitives express data values. In a bar chart, the heights of the bars typically represent the values that were captured for a given variable. Connotation, conversely, refers to all the supplementary ways in which a graphical element may provoke meaning. Viewers of a bar chart, for instance, are likely to view the data as a "discrete rather than a temporal trend", and line graphs "tend to evoke temporal interpretations".[575] It may be added that standardised graphics, such as bar charts and line charts connote quantification and objectivity, while the manually drawn visuals of Kerouac appear to belong more naturally within the realm of literary interpretation. For scholars who are sensitive to the aesthetic nature of graphics, Posavec's creative visualisations may bolster hermeneutic processes more cogently.

Visualisations, as was shown, can be dissected into their constituent meaningful components, and designers often have a wide range of options for each of these distinct aspects. Given a particular dataset, there is no predetermined way in which these values may be rendered graphically. Visualisations seldom represent data in a fully neutral manner. In response to Edward Tufte's axiom that graphics ought to "reveal data", Johanna Drucker stresses that data do not have an indisputable shape prior to their visualisation, and that "every graphic representation is a rhetorical device". Unlike scientific visualisations, which mimic an object or an environment whose visual characteristics are given a priori, diagrams that represent literary research data are largely designed during the data analysis phase. Creators of graphics often aim to convey particular ideas, and design their visualisation in such a way that they most effectively and most understandably express their predilections. In Ben Fry's description of the visualisation pipeline, the step that is labelled "refine" explicitly aims to further embellish or to improve the rhetorical effect of the graphics that are created initially.[576] By altering aesthetic aspects such as shape, colour and width, or, by adjusting ranges on the horizontal or on the vertical axes, for instance, researchers can actively influence the dimensions and the overall appearance of the patterns that emerge. Via such manipulations of

[574] <http://www.stefanieposavec.co.uk/> (10 March 2014)

[575] Jessica Hullman & Nicholas Diakopoulos, "Visualization Rhetoric: Framing Effects in Narrative Visualization", p. 2237.

[576] Ben Fry, *Visualizing Data*, p. 5.

visualisations, viewers may be guided into the direction of specific readings. Maureen Stone has noted that such manipulations may occasionally lead to misrepresentations of the data.[577] Kathleen Kerr explains, in a similar fashion, that visualisations can be based either on a formalist and rational approach, or on a social constructionist and post-rationalist approach. The former approach assumes that the characteristics of data can be shown in a manner that is unproblematically truthful, and that visualisations can represent the inherent shape of the data set in an objective manner. The social constructionist approach emphasises that "the right format is negotiated in relation to the needs of both producer and user".[578] Graphics are designed by scholars working in a particular theoretical tradition, using tools that are based on particular epistemological assumptions, and for the purpose of highlighting particular aspects in the data set. What is included and what is excluded in a visualisation depends for a large part on the idiosyncratic interests and the methodological conventions of scholars.

Whereas scholars can usually influence the appearance of the forms that are produced out of data, making an explicit claim about the domain that is depicted, beyond the mere patterns that are produced, is generally onerous. Data visualisations primarily depict patterns, and they generally fail to express sustained arguments. Jay Bolter explains that images, in general, are "designed to identify objects ... and situations ... rather than convey a discursive message".[579] Defending an argument involves a temporal or a linear progression in which initial premises are followed by statements that can be concluded from earlier assertions. While different images can evidently be placed in a sequence in order to construct a narrative, the precise relations between the ideas and the concepts that are depicted is often difficult to convey explicitly, without taking recourse to the verbal modality. Persuasive argumentation also demands the possibility to express evaluative statements, or the capacity to categorically sanction or to dismiss an idea. Images can depict objects or phenomena, but, since there are no codified visual means of expressing approval or dismissal, it is often difficult to give a direct and unambiguous expression to the attitude of the creator of the graph towards the objects that are depicted. Visualisations, in short, are often unable to present complicated forms of inductive reasoning, or to invalidate counterarguments.

Data visualisations, crucially, convey patterns rather than sustained arguments. Precisely because of the dearth of explicit arguments, however, viewers are often empowered to develop their own readings of the visual information, within the constraints that are set by the creator of the graphic. Sarah Jones emphasises that textual descriptions of a scholar's impression of a literary work habitually remain

[577] Maureen Stone, "Information Visualization: Challenge for the Humanities", pp. 46–47.

[578] Kathleen Kerr & Waqas Javen, "Visualization and Rhetoric: Key Concerns for Utilizing Big Data in Humanities Research", in *IEEE International Conference on Big Data*, (2013), p. 29.

[579] Jay Bolter, *Writing Space: The Computer, Hypertext, and the History of Writing*, p. 61.

confined to those aspects which were considered worthy of attention by the author. Readers of a text follow a discursive trajectory which was fully designed beforehand on the basis of the interests of the author. Data visualisations, by contrast, can grant viewers the liberty to concentrate on the aspects of their choice. Compared to reading a text, inspecting a chart or a diagram is an experience that "belongs more to the reader".[580] The ability to provide direct access to visual renditions of data may be considered an exponent of the broader development which is referred to by Van der Weel as the "deferral of the interpretative burden".[581] As digital research instruments continue to generate large quantities of raw data, scholars increasingly provide access to scholarly "semi-manufactures". Visualisations do not explicitly state why particular shapes are relevant or significant, and they lack an explanation of the causes that underlie the patterns that are shown, relocating the the task of "sense-making" to consumers of these resources.[582] As discussed, it would be incorrect to suggest that data visualisation provides access to uninterpreted data. The data values themselves often result from subjective methodological decisions, and their graphical renditions are likewise the outcome of a biased selection of visualisation techniques and of subsequent subjective refinements.

The suspension of explication complicates Jessop's suggestion that visualisations of data can serve as substitutes for textual publications. The question whether or not graphics can serve independently as legitimate products of scholarship can be connected to a broader discussion on the status of non-textual artefacts produced within digital humanities research. A scholarly engagement with computers generally results in a broad range of digital resources, such as databases, visualisations, software and scans. Many institutions that assess academic performance, including grant committees and tenure committees, frequently contend that the creation of such digital artefacts does not constitute scholarship in itself. In many cases, digital humanists can receive recognition for such digital products only when they conjointly publish a critical companion text. Mills Kelly has argued that genuine acts of scholarship ought to be "the result of original research". It must propagate "an argument of some sort", which needs to be "situated in a pre-existing conversation among scholars". [583] As data visualisations often lack such explicit arguments, it can be difficult to let them count as independent scholarly resources.

[580] Sarah Jones, "When Computers Read: Literary Analysis and Digital Technology", in: *Bulletin of the American Society of Information Science and Technology*, 38:4 (2012), p. 28.

[581] Adriaan van der Weel, "New Mediums: New Perspectives on Knowledge Production", in: Wido van Peurzen, Ernst Thoutenhoofd, & Adriaan van der Weel (eds.), *Text Comparison and Digital Creativity*, Leiden: Brill 2010, p. 263.

[582] In this respect, data visualisations are similar to digital editions which provide access to all extant witnesses of a text. If the resource does not privilege a particular variant, readers are forced to find their own paths through the materials.

[583] Mills Kelly, "Making Digital Scholarship Count", in: Dan Cohen & Tom Scheinfeldt (eds.), *Hacking the Academy: A Book Crowdsourced in One Week*, Michigan: MPublishing 2011, p. 33.

The notion that visualisations are incapable of expressing interpretation has been contested, however, by Johanna Drucker. She urges scholars to “find graphical means of expressing interpretative complexity”, and also notes that standardised visualisation models, such as bar charts and pie charts, are inadequate in this respect, largely because of their roots in statistics and in the empirical sciences.[584] The field of literary studies does not have a historical traditional of using visual displays, and digital humanists were consequently forced to borrow many tools and techniques from other disciplines, such as the social and the natural sciences. Drucker emphasises that such inherited visualisation techniques inappropriately follow the epistemological assumptions of disciplines that assume that properties of phenomena exist independently from the observer. Conventional graphical displays which represent values on a Cartesian coordinate system, or which represent values as distinct bars with equivalent sizes are considered inappropriate for the representation of the results of humanities research. The concepts and phenomena that scholars engage with critically are frequently highly complex, and they typically resist a reduction to a finite set of sharply separate categories. Specific phenomena, such as the experience of time in Woolf’s novel *Miss Dalloway*, for instance, cannot be represented simply though a linear timeline. Drucker has developed experimental diagrams herself which draw clear attention to the interpretative and subjective nature of humanistic enquiry, by using fluid borders in between categories or by replacing single broad categories with a multitude of smaller overlapping more flexible categories. Instead of Cartesian grids, Drucker proposes axes in which the temporal dimension is more transient. The bar charts she proposes do not display values in mutually exclusive categories, but values which can be in different categories simultaneously and values which can flow unstintingly outside of predefined categories.

Drucker’s diagrams may be viewed as a criticism, in a graphical form, of an oversimplified approach towards quantification in the humanities. Indeed, the reductions and the abstractions that underlie data creation are not always warranted. Properties cannot always be described via a single value and texts do not always fit neatly within predefined categories. Drucker proposes that the concept of data, approached in a reductive manner, must be replaced with the notion of interpretation, being the researcher’s subjective experience of the properties of aesthetic works. Evidently, such idiosyncratic impressions do not follow mathematical principles, and, consequently, these cannot be visualised using standardised statistical packages.

While Drucker’s experimental diagrams rightly attest to the problematic character of humanistic data, the suggestion that visualisations ought to express the full complexity of the phenomena that are studied can be challenged. Drucker argues that visualisations created for humanities research ought to represent

[584] Johanna Drucker, “Humanities Approaches to Graphical Display”, s. 50.

subjective interpretations of phenomena rather than quantitative data which can be manipulated mathematically, but such renditions of interpretations are often difficult to interpret in themselves. Graphs are typically produced to enhance cognition, and to impart information about a specific scholarly domain. Drucker's humanistic visualisations, consisting of fluid categories and non-linear timelines, seem difficult to decode without an additional textual explanation. The graphs do organise specific components in a spatial arrangement, but it is unclear if these specific positions also represent a particular meaning. The diagrams, much like the works that are depicted, become resources that need to be interpreted. Since these visualisations are not based on a formal logic or a known code, and given the limited expressivity of the graphical modality, the viewer cannot know whether or not a particular interpretation is correct until a textual explanation is given. As the precise statements about the phenomena that are studied are often unclear, the diagrams primary emphasise the notion that the creation of data requires nuance and flexibility. When graphical elements are used to communicate a critical scrutiny of the humanistic methodology, as in Drucker's proposal, this imposes an illocutionary weight on these graphs they can barely carry.

An additional difficulty inherent in graphs which represent the interpretation of phenomena is that they frustrate the emergence of new discoveries. Data visualisations are often created to marshal surprises. They may expose unexpected correlations or conspicuous disassociations, which can in turn galvanise new readings of the texts that are rendered graphically. The visualisations proposed by Drucker are steered wholly by existing interpretations, which, as can be assumed, were produced in a non-digital context. Evidently, the embossed presence of one scholar's interpretation discourages and restrains new exploratory hermeneutic analyses. Drucker's diagrams indeed usefully stress the convolution of particular humanistic phenomena, but, as the diagrams are ostensibly not based on logic, they remain difficult to interpret. Mathematics and statistics, by contrast, imply a standardised method for encoding and decoding graphical properties, and it seems judicious to continue to create diagrams which adhere to basic quantitative principles. Visual literacy, in many cases, demands graphical displays based on data which can be manipulated mathematically. Discussion about the complexity of the methodology, and, more specifically, about the difficulties that inhere in the definition and the application of descriptive categories, can arguably be expressed more effectively via the textual modality.

Visualisations, in short, communicate characteristics of quantitative data via particular constellations of graphical primitives. They may present data at different analytic levels and they may be static or interactive. While visualisations are less adept at communicating complex scholarly arguments, visualisations often stimulate their viewers to formulate interpretative claims on the basis of patterns. Gibbs and Owens stress that work with data in the humanities "can be exploratory and deliberately without the mathematical rigour that social scientists must use to support their epistemological claims".[585] The capacity of data analyses to generate new ideas is generally more important than their ability to serve as conclusive evidence for particular statements, let alone than their ability to make such statements themselves. Visualisations used in the service of literary criticism are valuable mostly because of the fact that they can foster a hermeneutic engagement with the objects that are represented.

9.4. Interpretation

In *Simulacra and Simulation*, Jean Baudrillard has argued that postmodern society has become deeply reliant on models and on simulations. He uses the term "simulacrum" to refer to "models of a real without origin or reality".[586] While representation is based on the principle that the sign is a counterfeit of an original, simulacra are models of objects or phenomena which are fully contrived. They establish a "hyperreality". Visualisations may likewise be viewed as simulacra. Helen Westgeest observes that, within the natural sciences, graphics are often created to give objects which are invisible in reality a perceivable presence. Researchers can visualise imperceptible phenomena by means of models, which can be described as "visual constructions".[587] Visualisations of data about literary texts likewise fabricate shapes that do not have an exact match in an objective reality. Data sets are evanescent and shapeless in themselves, and the process of visualisation fundamentally consists of inventing a physical layout for these data. Graphical displays introduce new forms, but they are produced to reveal characteristics of the originals which are represented. Their aim is mostly to produce an innovative perspective, and to allow for novel and compelling interpretations of these texts.

For interpretation to be possible, viewers initially ought to be able to decode the visual patterns that are observed into descriptive statements about the objects that are represented. They need to be able to formulate a set of statements which appear to be justified on the basis of what is shown in the visualisation. Stephen Few stresses that a graphic is "only successful to the degree that it encodes information in a manner that our eyes can discern and our brains can understand".[588] The value of visualisation tools may be assessed by evaluating how effectively they

[585] Frederick W. Gibbs & Trevor J. Owens, "The Hermeneutics of Data and Historical Writing", in: Kristen Nawrotzki & Jack Dougherty (eds.), *Writing History in the Digital Age*, MIT Press 2013, n.pag.

[586] Jean Baudrillard, *Simulacra and Simulation* (Ann Arbor: University of Michigan Press 1994), p. 166.

[587] Helen Westgeest, "Visualizing Research and Visual Communication of Research", in: Helen Westgeest (ed.), *Making Research Visible to the World*, Amstelveen: Canon Foundation in Europe 2010, p. 12.

[588] Stephen Few, "Data Visualization for Human Perception", n.pag.

convey information. When data sets are voluminous and heterogeneous, it can evidently be taxing to represent the values in a manner that is both legible and meaningful. A comprehension of the information that is represented may additionally be impeded when viewers lack an understanding of the type of processing the data have undergone. One example of a tool which may clog cognition is Voyant Knots. The tool represents the corpus as a collection of twisted lines, and the "extent to which lines overlap indicates the level of correspondence or linkage between the terms".[589] Without a thorough understanding of the algorithm that underpins the visualisation, it is difficult to translate the visual patterns that can be generated in the tool to a univocal descriptive statement about the corpus that is depicted.

In the context of literary criticism, visualisations of data about texts can be valuable if these can spur a hermeneutic engagement. Interpretation, on a first level, entails an understanding of the central meaning or of the central theme of the text. If the term "theme" is understood narrowly as the direct denotation or as the literal "aboutness" of the text, explorations of themes may potentially be based on displays of significant or distinctive vocabulary. This approach is followed for instance, in Matthew Jockers' book *Macroanalysis*.[590] As was discussed in Chapter 5, poetic texts rarely express their abstract literary themes directly. Themes can, in many cases, be found exclusively via close reading. In writing that uses figurative language, the semantic context is generally too complex and too unpredictable for current semantic taggers or for topic modelling algorithms. It is cumbersome, for this reason, to devise visualisations which enable scholars to read the literary themes a collection of texts directly.

Whereas visualisations often fail to elucidate the immediate thematic concerns of a poem, graphical resources can help scholars to investigate the manner in which the language of a text contributes to the production of its central meaning. The aim of conventional close reading is often to describe the relationship between the text's form and content. Data which are produced automatically typically focus on the language of the literary object, and operations such as filtering, clustering, sorting or distribution, performed at the level of the corpus as a whole, can often expose trends or regularities in an author's use of literary devices or of linguistic constructions.

As was discussed in Chapter 4, the various functionalities which are provided by text analysis tools can be classified by making use of Unsworth's concept of the scholarly primitives. Visualisation tools primarily provide support for "comparison" and for "discovery". Graphics which display correlations, clusters or distributions centrally aim to expose differences and similarities between the texts within the collection and, as such, they centrally entail a comparison of texts.

[589] <http://docs.voyant-tools.org/tools/knots/> (2 June 2015)

[590] Matthew Jockers, *Macroanalysis : Digital Methods and Literary History*, p. 118.

Comparison initially discloses patterns within a collection of texts, but it may also lead to the discovery of a limited set of documents which are distinctive within the context of the collection. Visualisations of text corpora primarily support interpretation through their capacity to inspire more targeted forms of close reading. Visualisations of the formal features of texts often raise particular questions, and the answers to these questions can often be found solely by rereading the separate texts that shape the diagram. The formal properties of texts can be accounted for by relating these to other factors such as the genre, the historical context or the overall thematic concerns.

While the close reading that follows from discovery can be performed on the basis of the original linear texts, such analyses at the level of poems or at the level of verse lines can also be supported effectively via visualisations which represent data about individual poems. Visualisations at the micro-level have been developed within a number of projects. The Mandala browser which was developed by Dobson, Ruecker, Gabriele and Sinclair, primarily offers support for the clustering of the smaller units of a texts. The browser demands XML-encoded texts as input. On the basis of the encoding, the text is divided into its constituent structural components, such as paragraphs, verse lines, or, in the case of drama, speeches. Having created these smaller units, users can define search terms within the text's interface which can attract or repel specific fragments. In this way, specific groups can be created of text fragments which contain or which lack specific search terms. As the Mandala Browser transforms the single linear text into an assemblage of movable units, Brown et al. note that the application stimulates a traversal of "a major interface boundary". The verbal text is abandoned "in favour of working with configurations of coloured dots",[591] and scholars can focus predominantly on the formation and the exploration of patterns.

Visualisations of individual texts can similarly be produced within *Myopia* and the *Poem Visualizer*. In *Myopia*, data about literary devices initially need to be supplied manually via a customised set of TEI elements. The program can subsequently exploit these data to generate abstracted renditions of the encoded texts. Myopia displays the individual words of the poems as blocks, and it can be specified that the fill colours of these blocks must be determined by the occurrences of literary devices. The tool can illuminate the ways in which textual phenomena such as alliteration, rhyme or consonances are distributed over the various lines.[592] *Poem Viewer*, which was developed by A. Abdul-Rahman et al., likewise shows the full verse lines, combined with connectors at the end of each line which indicate the lines that rhyme. A small diagram is shown, furthermore, above each vowel in the text, to clarify the positions of phonetic articulations. Such

[591] Susan Brown et al., "Reading Orlando with the Mandala Browser: A Case Study in Algorithmic Criticism via Experimental Visualization", in: *Digital Studies / Le champ numérique*, 2:1, n.pag.

[592] Manish Chaturvedi et al., "Myopia: A Visualization Tool in Support of Close Reading", n.pag.

graphics effectively enable scholars to study developments in vowel positions and particular sonic recurrences.[593]

Unlike the Mandala browser, *Myopia* and the *Poem Viewer* display both the original linear texts and the data that are available about these texts. In a sense, these types of visualisations blend the textual modality with the graphical modality. They blur the distinction between reading texts via typography and viewing texts via visualisations. An important objective of visualisations is to establish a condensation, and to provide a succinct expression of the available data. In the visualisations that are created in *Myopia* and in the *Poem Viewer*, however, the graph is as extensive as the text itself. The graphics also display a degree of linearity, since the order in which the data elements are shown corresponds directly to the linear order in which these aspects occur in the original. Graphics which display the occurrences of literary devices within the same logical structure that is used in the original linear text can effectively illustrate the craftsmanship of a poet. Poetic effects frequently consist of specific arrangements of sounds or of words, and the exact nature of correspondences and recurrences can often be illuminated adequately by visualising occurrences of such devices at the level of individual verse lines.

Such visualisations at the micro-level, in which occurrences of literary devices are indicated within their original logical structure can similarly be created using the structured annotations that were created for the central case study of this thesis. Figure 9.1 clarifies the distribution of perfect rhyme and of alliteration in MacNeice's poem "Selva Oscura". Occurrences of perfect rhyme are shown in red and instances of alliteration are shown in blue. In some cases, the final phoneme sequences which produce the rhyme at the end of the verse lines also appear in the interior of verse lines. Such forms of internal rhyme have also been indicated using a red colour. The poem has a fairly regular structure, consisting of four stanzas of five lines each. In each stanza, only four lines rhyme. This formal structure, in which one line is left companionless, mirrors the loneliness of the poem and the central idea that a "house can be haunted by those who were never there / If there was where they were missed" (ll. 1-2). The poem synchronously makes a commanding use of alliteration. In many of the verse lines, and most frequently in the last two stanzas, the word which produces the end rhyme also alliterates with one or more other words on that same line. This is the case in the lines "One sudden shaft of light from the hidden sky" and "Perhaps suddenly too I strike a clearing and see". The alignment of rhyme and alliteration, and the harmonious ring that it gives to these lines, effectively braces the experience, captured in the final stanza, that

[593] A. Abdul-Rahman et al., "Rule-based Visual Mappings – with a Case Study on Poetry Visualization", in: B. Preim, P. Rheingans, & H. Theisel (eds.), *Eurographics Conference on Visualization (EuroVis)*, (2013), p. 8.

the solitude can also be ended. In the final lines, the speaker sees that the door of the isolated house “swings open and a hand / Beckons to all the life my days allow”.

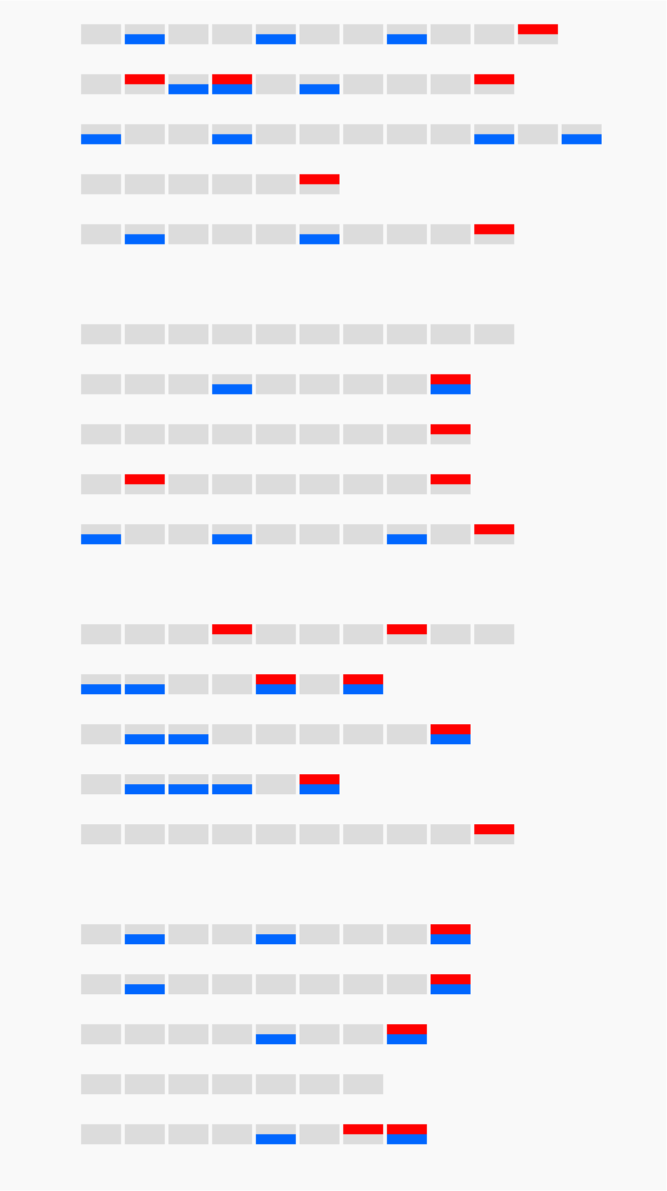


Figure 9.1. Visualisation of perfect rhyme and of alliteration in “Selva Oscura”

Figure 9.2 clarifies the distribution of rhyming sounds in the "The Sunlight on the Garden". To create this diagram, the rhyming sounds that occur in the poem have been assigned a unique code. All words have been analysed, so that occurrences can be identified of both final rhyme and internal rhyme. The diagram effectively displays both the elegant rhyming scheme and the adept use of internal rhyme. A degree of editing was necessary, nevertheless, to ensure that the regularities in the rhyming patterns could be shown effectively. It was found, for example, that the regularity in the end rhymes could be shown most clearly if the verse lines were aligned to the right.

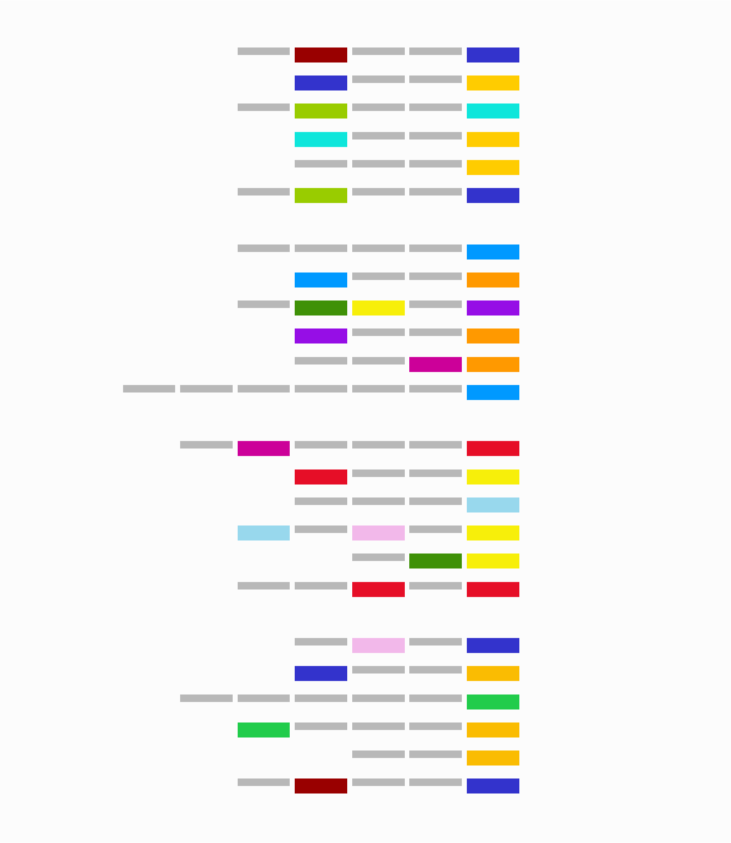


Figure 9.2. Rhyming sounds in "The Sunlight on the Garden"

Figure 9.3 is a similar representation of "The Glacier". As is the case in figure 9.2, all the final phoneme sequences that occur more than once, both at the end and in the interior of a verse line, have been associated with a unique colour code. The colours have been assigned randomly. As can be seen in figure 9.3, the first section contains a large number of repeated sounds. Phoneme sequences are repeated, for instance, in the words “who”, “through”, “two” and “you”, in “climb” and “fine” and in “they” and “day”. In this poem, the use of rhyming sounds is clearly supportive of the text’s central idea. “The Glacier” consists of two separate sections. The opening section focuses on the hectic nature of modern life, and it conveys the central idea that the traffic on the city streets moves so quickly that it paradoxically seems static and petrified. The closing section expresses a longing for a quieter alternative. The cascade of repeated rhyming sounds in the first half of the poem, in a sense, mimics the glacier that is depicted. The second section of the poem contains a smaller number of rhyming sounds and the lower half of the diagram consequently looks calmer. The calmness and the stasis is also underscored by the rime riche in “where” and “ware”. While patterns such as these are difficult to see in the regular verbal structure, such unequal distributions can be shown efficiently via visualisation.

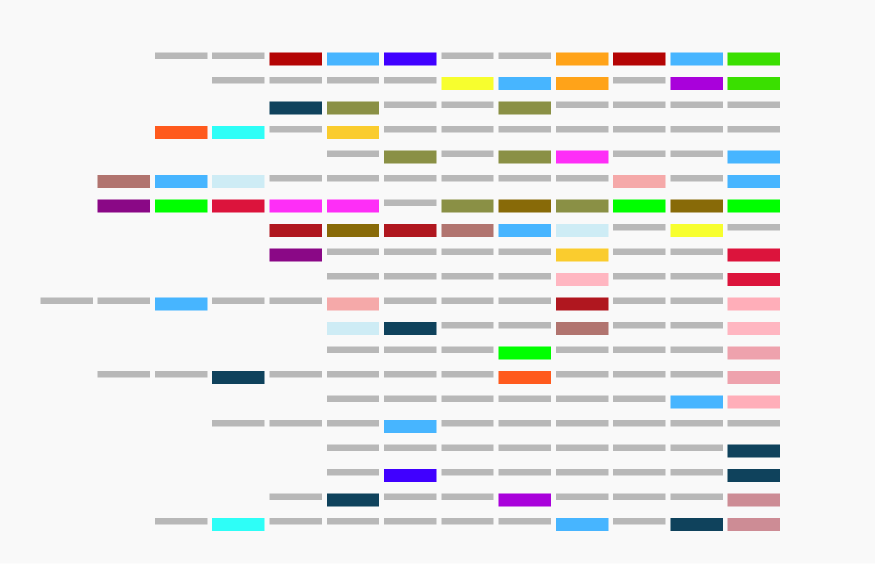


Figure 9.3. Rhyming sounds in “The Glacier”

Visualisations at a micro-level can also clarify the texture of the poem. The term *texture* generally refers to the manner in which the sounds that occur in a poem are combined into patterns. The sounds in vowels and in fricative or aspirated consonants are mostly soft, while plosives are often experiences as harsh. Figure 9.4 is a representation of the texture of “The Glacier”. To produce this graphic, each line has been divided into its separate syllables. Syllables containing plosives have been given a dark red colour, while the syllables which contain softer sounds have been assigned a light yellow colour. On the basis of this visualisation, the softness or the harshness of the sounds in the different poetic lines can be studied directly. It can be seen that the first section of the poem uses many harsh sounds. This is especially the cases for lines 6 and 12 of the poem ("That you cannot catch the fraction of a chink between the two" and "Cannot bear to watch that catafalque creep down"). The second part, which expresses a longing for quietude, also uses much softer sounds. Line 18, for instance ("Eyes appraise the glazen life of Majolica ware") mostly contains nasal and lateral consonants.

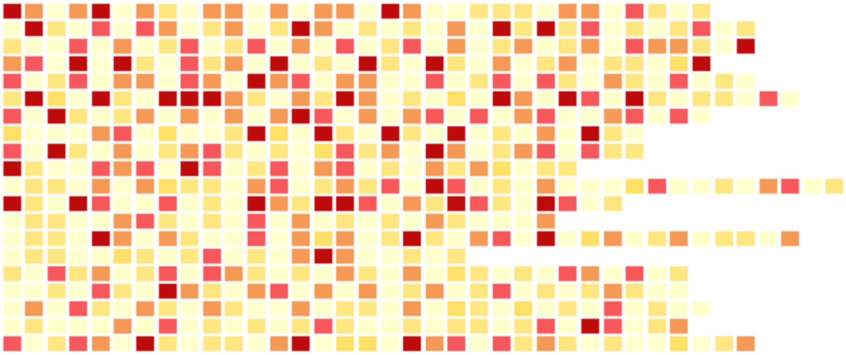


Figure 9.4. Visualisation of the texture of "The Glacier"

A central objective of visualisation, and more, generally, of data analyses, is to arrive at patterns which can stimulate interpretation via more focused forms of close reading. Graphical representations of individual texts can elucidate specific patterns in the use of literary devices, and such patterns can encourage scholars to explore whether or not these formal features support and reinforce the meaning of the text. The aim to produce such interpretable patterns simultaneously implies a predicament. Texts can be processed and visualised in an endless number of ways, and it is often difficult to know beforehand if particular manipulations can actually result in patterns that can foster a hermeneutic engagement. Some manipulations may open up new interpretative possibilities, while other manipulations may prove unproductive. As the precise effects of statistical operations are often unforeseen, the discovery of patterns typically contains an erratic and an aleatory element. The production of interpretable graphics often demands various cycles of trials and of errors.

Literary texts are invariably highly complex and multifaceted phenomena, and through the process of visualisation, researchers can reduce such complexity. Graphic renditions of quantitative data mostly aim to clarify the main characteristics of multivariate data collections by rendering these in a succinct format. Such abstractions or summarisations unavoidably imply a loss of information, as they are created by privileging a limited set of dimensions at the expense of certain other dimensions. There can subsequently be a more concentrated examination of the dimensions that remain. Visualisations are generally abstractions which simplify complex data sets and which obscure much of the ambiguities and uncertainties that may interfere in these data.

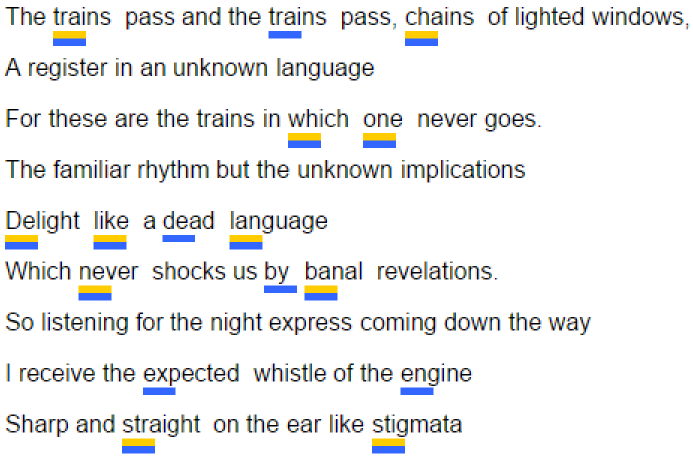
While such simplifications may be useful, literary scholars may also benefit from methods which can expose the nature of such reductions and which can effectively indicate the ambiguities and the contradictions that exist within the objects that are studied. Kouw et al. stress, more generally, that ambiguity and uncertainty can frequently form a source of new knowledge. Ambiguities can serve as a disruptive force which can illuminate the ways in which prevailing conceptualisations misrepresent an actual situation. When scholars are attentive to uncertainty, this "opens up potentials otherwise veiled", as problems of uncertainty and ambiguity may provoke new types of solutions. While computer applications are often concerned with disambiguation and with reducing complexity and irregularity, literary informatics can also benefit from applications that support processes of reambiguation, and from visualisations that can effectively demonstrate that texts are more complicated and more varied than was originally assumed. Digital methods can stimulate interpretation by generating compelling quandaries which may force scholars to read texts in a different light. Scholars do not necessarily aim to solve all the difficulties that inhere in a poem, and they are frequently intent on demonstrating that texts can yield to many different, and potentially contrastive, readings.

At present, however, there are no unequivocal rules or guidelines for creating visualisations that can productively expose complexity and ambiguity. In natural languages, speakers can often use qualified language to indicate that specific assertions lack convincing support. Levels of ambiguity or of uncertainty can be made explicit, for instance, via the choice of specific verbs and adverbs. When data sets include ambiguous data, it is pivotal, analogously, to qualify these visualisations. Leland Wilkinson stresses that data sets may be incomplete, indeterminate, biased or downright erroneous, and graphs which do not explicitly indicate the presence of uncertain or faulty values are obviously deceptive. Wilkinson notes that when a data set contains systematic indications of the confidence levels of variables these can used in a variety of ways. Such quantifications of the levels of uncertainty can be used to determine the degree of transparency, the textures or the shapes of the icons and of glyphs which are used in a visualisation. Such visual aids may "guide, qualify, or soften our judgments of uncertain data".[594]

In studies that concentrate on data about literary texts, three causes of ambiguity can be distinguished. A first form of ambiguity is generated by a conflict between the connotations of literary devices and the thematic concerns of the texts in which these devices are used. A number of such cases of ambiguity have been discussed in Chapter 8. Poetry written in light verse is generally experienced as

[594] Leland Wilkinson, *The Grammar of Graphics*, p. 450.

buoyant or as cheerful, but, as was shown, MacNeice often used this verse form in poems which address morose and gloomy subjects. Visualisations which juxtapose phenomena which seem incompatible or contrastive can often pose stimulating quandaries.



The trains pass and the trains pass, chains of lighted windows,
A register in an unknown language
For these are the trains in which one never goes.
The familiar rhythm but the unknown implications
Delight like a dead language
Which never shocks us by banal revelations.
So listening for the night express coming down the way
I receive the expected whistle of the engine
Sharp and straight on the ear like stigmata

Figure 9.5. Two alternative methods for detecting alliteration

The fact that many of the data have been generated via algorithms forms a second cause of ambiguity. Using a term proposed by Mons and Velterop, the annotations that have been produced automatically may be referred to as “hypothetical data”.[595] The algorithms that have been used to create these data are critically based on a theory about the manner in which particular literary phenomena may be detected. These data remain hypothetical until they have been vetted by human scholars. If manual interventions in the results of the algorithms are registered in metadata, visualisations may also represent the difference between hypothetical data and data which are endorsed manually. Alternatively, the hypothetical nature of data about literary texts may potentially be emphasised by visualising the results of different implementations of algorithms. In Figure 9.5, the colours beneath the words indicate the implications of two different approaches towards the recognition of alliteration in MacNeice’s poem “A Contact”. The first algorithm, whose results are shown in yellow, considers repetitions of consonants in stressed syllables. The second algorithm asserts that all repeated sounds at the beginning of both stressed and unstressed syllables can produce alliteration. Visualisations such as these effectively demonstrate the notion that idiosyncratic

[595] Barend Mons & Jan Velterop, “Nano-Publication in the E-Science Era”, p. 8.

decisions about the logic that is used within an algorithm can have direct implications on the eventual visualisations. They may simultaneously expose cases in which two different implementations of an algorithm both result in relevant cases, underscoring the complexity both of the heuristic activity that is modelled and of the text that is analysed.

The ambiguity of data sets can also be connected, thirdly, to the inevitable subjectivity of observations. Even when specific assertions which were generated algorithmically are approved by one particular scholar, other scholars may still disagree fervidly with these observations. In the case of subjective, or observer-dependent, observations, it is generally impossible to quantify the level of certainty. Whereas visualisations typically reflect one particular author's preferred manner of simplifying, static graphs misleadingly evoke a sense of closure, and they consequently acquire an aura of objectivity. Rieder and Röhle observe that there is a tendency among viewers to automatically accept visual information as evidence.[596] While arguments which are presented in a textual resource often disclose the larger process that has been followed to arrive at a particular conclusion, static graphics often lack information about their genesis. As it is difficult for viewers to identify the potential errors that have been made in the production of the static visual, it is correspondingly difficult to refute the findings that are presented.

In the digital realm, the seeming closure of data visualisations can be counteracted through the construction of interactive and elastic graphs. When other scholars are empowered to alter the parameters that govern the graphical rendition, this may effectively stress the notion that diagrams are also arbitrary and contingent. Interactive diagrams can illustrate the notion that different points of view also result in different diagrams. Interactivity strongly weakens a scholar's ability to communicate fixed and authoritative statements about texts, however, adding to the deferral of the interpretative burden. Interactive visualisations primarily help scholars explore the characteristics or a data set, to assess some of the assumptions that underlie the data, and to form new ideas about the objects that are depicted.

9.5. Conclusion

Louis MacNeice viewed verbal communication as a superior mode of expression. In his poem "To Posterity", MacNeice expresses his anxiety that this manner of communicating may eventually be replaced by "other, less difficult, media". He fears that, when experiences are no longer "framed in words", birds will "be always wingless birds". In digital humanities research, scholars increasingly make use of visualisations, and, these graphic displays, to a large extent, are indeed "less

[596] Bernhard Rieder & Theo Rohle, "Digital Methods: Five Challenges", in: David Berry (ed.), *Understanding Digital Humaniities*, Basingstoke: Palgrave Macmillan 2012, p. 74.

difficult", in the sense that they offer less refined possibilities for the communication of complex ideas. It is not feasible, for instance, to convey a sustained argument. Although visualisations are generally created for a particular purpose and from a particular theoretical perspective, they do not explicitly sanction or reject specific views. Furthermore, there are currently no widely accepted conventions for the visual expression of subjectivity and ambiguity. For this reason, it does not seem probable that visualisations will fully supplant discursive and argumentative texts. A visualisation is primarily a rhetorical device which can support an argument, and not an argumentation in itself.

Graphical displays do support a new form of studying texts, however. Data visualisations form compelling examples of the form of research which Jerome McGann described as "performative" and "deformative criticism". McGann surmises that computational analyses may reinforce hermeneutic processes via the creation of transformed versions of texts. Deformative operations, such as the elimination of specific word types, a reversal of the order in which lines are presented or the creation of frequency lists can expose unexpected perspectives on familiar texts, and these new vistas may subsequently spawn new ideas. Aspects such as "the structural forms of words, phrases, and higher morphemic and phonemic units" are often viewed as "preinterpretative and precritical",[597] but a rearrangement of these components can fruitfully "release or expose the poem's possibilities of meaning".[598] Visualisation derived from data about the literary techniques can be likewise valuable for literary critics, as they allow for a highly systematic scrutiny of the language that is used within a text. Works of literature are often very rich repositories of ideas and of sonic patterns. Using the graphical modality, the nature of text corpora may be illuminated via the production of distribution graphs, via filtering on the basis of secondary data, and via the creation of clusters. These basic operations result in patterns which may provoke specific ideas and interpretations.

Importantly, visualisations can be created at any level of aggregation. Next to visualisations of corpora in their entirety, scholars can also produce abstracted renditions of individual text fragments. Visualisations of separate texts may help to investigate the more precise ways in which words, sounds and grammatical categories collaborate to produce a meaning. The macro-level and the micro-level patently complement each other, and an exclusive focus either on the whole or on the units seems incomplete. A recognition of the deficiency of such a singular and unadaptable approach is also manifest in the appellation of one particular visualisation tool which can provide support for the process of close reading, namely, the Myopia application. The name of the tool notably implies an inability to perceive objects which are far away. It may be argued that, if close reading is

[597] Jerome McGann, *Radiant Textuality: Literature after the World Wide Web*, p. 115.
[598] Ibid., p. 108.

myopic, distant reading is a hyperopic form of engagement. In the latter form of reading, the focus is predominantly on the broader patterns, at the expense of particularities of individual texts. Unlike healthy eyes, eyes affected by myopia or by hyperopia are deprived of the capacity to focus variably. In the case of such defects, technology such as glasses or contact lenses may intervene to restore the ability to focus both on details which are nearby and on objects which are distant. Literary critics are arguably myopic by default, but they can similarly use technology to study the phenomena which can only be observed from a distance. Both perspectives are important for a solid understanding of literary works. Digital methods, importantly, enable scholars to switch between myopic and hyperopic reading, and to concentrate both on patterns and on particulars.

Chapter 10

Conclusion

10.1. Introduction

This thesis has established a number of salient characteristics of algorithmic criticism, and has explored some of the ways in which algorithmic processing of textual data may expand or curtail interpretative possibilities. It has been explained that machine reading entails a consistent and context-independent form of processing which can result in abstracted renditions of individual texts and of text collections in their entirety. As was demonstrated in the discussion of the central case study, the statistical resources and the visualisations that can be created through algorithmic analyses can often lead to fresh perspectives and to new ideas about literary works. The properties of algorithmic criticism that have been discussed in the previous chapters are likely to affect the broader field of literary studies in a variety of ways. Section 10.2 ruminates on some of the fundamental ways in which algorithmic criticism differs from conventional criticism. Four important differences can be identified: (1) it places a greater emphasis on practical work; (2) it lead to different forms of scholarly output; (3) it results in new ways of discovering texts with noteworthy characteristics; and (4) it can supply different types of arguments to support scholarly claims. It is important to stress, nonetheless, that there are also a number of important continuities: (1) the ways in which digital technologies are implemented are strongly marked by a subjectivity; (2) it uses both inductive and deductive methods; and (3) scholars continue to bear the responsibility to evaluate whether or not the methodology adequately supports the discipline's central scholarly objectives. The main similarities are discussed in section 10.3.

Through this focus on the various ways in which technology affects literary scholarship, this study aims to answer to David Berry's and Alan Liu's calls for a more critical mode of digital humanities research. Such a critical approach ought to be attentive to "the digital component of the digital humanities in the light of its medium specificity, as a way of thinking about how medial changes produce epistemic changes".[599] By studying the various ways in which computational methods may affect existing conceptualisations of knowledge, this thesis also aimed to make a contribution to the emerging scholarly field of software studies. Basset explains

[599] David M. Berry, "Introduction", p. 4.

that software studies “turns what were once understood as the supporting dimensions of digital culture to the fore, and takes as the central problematic the cultural operations of software, and in particular the relationship between language and code”.[600] More broadly, software studies seeks to identify the essential characteristics of software systems and critically examines their social and political effects. Since, as is stressed by David Berry, “certain social formations are actualized through crystallization in computer code”,[601] it appears reasonable to assume that the use of computational methods can have important implications for scholarly practices within literary research. In evaluating the merit of using technology, it is crucial to consider whether or not new technologies can veritably enable scholars to ask new questions and to produce new forms of knowledge.

10.2. Changes

10.2.1. Practical work

An important difference between conventional criticism and algorithmic criticism is that the latter form of scholarship invariably demands work of a practical nature. Such practical work may entail, more specifically, the construction of a text archive, the preparation of a corpus of encoded texts, or the development or further refinement of tools for the analysis of digital materials. Scholars who aim to study texts via digital methods often face the difficulty that the sources they are interested in are not yet available in an authoritative machine-readable form. Digital scholarship, furthermore, depends crucially on tools with which these sources can be analysed. Since the analytic tools that are publicly available often concentrate on generic functions, or on collections of a limited size, they may not be suitable for differently focused research questions. Digital humanists often need to invest substantial amounts of time and intellectual efforts into the development of resources and instruments themselves, often in close collaboration with computer scientists or librarians. Digital humanities research demands two distinct classes of activities, which Jerome McGann characterises as a bifurcation between “conceptual undertakings (gnosis)” and concrete “constructions (poeisis)”.[602]

Despite the fact that the creation of tools and resources is often very labour-intensive, the development of such digital artefacts is not always recognised fully as a legitimate form of humanistic scholarship. It is often difficult for scholars to make the results of digital work count in assessments of scholarly productivity, as these traditionally privilege textual publications. Practical work is often viewed as a mere preparatory activity, necessary as a support for the more critical analysis that

[600] Caroline Basset, “Canonicalism and the Computational Turn”, p. 119.

[601] David Berry, *Critical Theory and the Digital* (London: Bloomsbury Publishing, 2014), p. 83.

[602] Jerome McGann, *Radiant Textuality: Literature after the World Wide Web*, p. 83.

take place at a later stage. Such a stance is misguided, however, as digital resources and tools generally demand critical analysis and intellectual exertions in themselves. Projects which aim to create scholarly digital resources frequently face a plethora of fundamental challenges which, in many cases, may only be addressed through a reference to more fundamental theoretical frameworks or concepts. Applying TEI, for instance, requires a deep understanding of the material that is encoded, and of the overall scholarly benefits that may be reaped from such editorial and critical interventions. Scholars who construct tools often need to take decisions about vocabulary, about the user interface or about the logic of algorithms, and such judgements are invariably based on theoretical assumptions.

While practical work often demands the construction of a prior theoretical framework, praxis can conversely lead to novel theoretical insights. McCarty stresses that the act of building also has epistemological value in itself. Modelling is "the continual process of coming to know by manipulating things".[603] Eventually, applications ought to function unobtrusively, but, before a tool can attain such a state of translucency, there is mostly a phase during which developers and adopters still question whether or not the tool can reliably and effectively be used to answer a question. When the algorithms that are implemented in a tool are applied to a corpus, this often exposes precisely those points on which the theoretical modelling misrepresents the actual situation. Such technical exigencies may necessitate a reconsideration of the logic that underlies a tool. The version that eventually emerges from the various alpha and beta versions may be seen as the conclusion that is drawn from these experiences. In this way, the development of an instrument contributes to a fundamental understanding of the nature of the task. Experimentation may reveal that particular aspects cannot be mapped directly to the strict information structure that is imposed by the computer, and creating a model is often "useful for isolating what gets lost when we try to specify the unspecifiable".[604] Julia Flanders views digital scholarship similarly as a form of translation. To be able to process artefacts digitally, parts of the existing discourse about cultural objects need to be converted to statements in a highly rigid and formalised language, and this often effectuates an estrangement. Paradoxically, the disunity between these different modes of expression can also be productive, as the assiduous work that is often needed to create the model invariably leads to an improved understanding of the activity that is modelled.[605]

In a more traditional form of research, the utility of computational methods would be investigated via an examination of the works of literary theorists, and by speculating subsequently about the obstacles that could arise if some of these critical activities are automated. An approach which fully evades practical work,

[603] Willard McCarty, *Humanities Computing*, p. 28.

[604] Ibid., p. 25.

[605] Julia Flanders, "The Productive Unease of 21st-Century Digital Scholarship", pp. 13–14.

and which is exclusively theoretical in nature is inadequate in studies which focus on the impact of the digital medium. Johanna Drucker emphasises that “abstract theory and critiques of the foundation of textuality in the terms of older philosophies” are in themselves insufficient to explain the type of knowledge that can be produced by digital tools.[606] To develop an understanding of the digital medium, an active involvement with the digital medium is indispensable. Such hands-on work is necessary because of the tacit nature of computing skills. In *Personal Knowledge*, Michael Polanyi argues that a proficiency in a practical skill entails a “tacit and passionate contribution of the person knowing what is being known”.[607] Tacit knowledge cannot be transferred via writing, and needs to be acquired in a practical setting and via experience. An understanding of the possibilities and the limitations of computer-based scholarship crucially demands an active engagement with coding. Algorithm-based analyses of textual materials frequently produce results that could not easily have been predicted or envisaged on the basis of theory alone.

Practical work enables scholars to produce knowledge about the methodology of the field. Text analysis tools generally advance an argument, often implicitly, about the textual aspects that are of relevance, about the manner in which these aspects can be recognised, and about the manner in which these aspects, once quantified, can be further processed. The development of a software tool for the automated discovery of literary allusion, for instance, demands tasks which are very similar to the type of work that would be needed for authoring a discursive scholarly text about the general nature of literary allusions. The construction of tools demands a precise definition of terminology and a clear hypothesis about the manner in which the phenomenon can be identified.

Next to providing support for the actual analysis and interpretation of literary works, algorithmic criticism additionally aims to evaluate whether or not innovations in the field of language technology can usefully be applied to study questions of literary criticism. The nature of the practical work that is performed within literary informatics is often very dynamic, as algorithms for the exploration of texts evolve incessantly. Computer science continues to annex territories in areas which were previously considered impermeable. Literary scholars with an interest in computing continually need to remain abreast of technological advances, and need to adjust their understanding of the type of data that can be produced about texts. More pertinently, they must also evaluate, on a more fundamental level, whether or not such technical innovations can genuinely extend the possibilities for understanding the value and the meaning of literary texts.

[606] Johanna Drucker, “Theory as Praxis: The Poetics of Electronic Textuality”, in: *Digital Poetics*, Tuscaloosa: University of Alabama Press 2002, p. 683.

[607] Michael Polanyi, *Personal Knowledge Towards a Post-Critical Philosophy.* (Chicago: University of Chicago Press 1958), p. 329.

Unlike literary interpretations, the knowledge that is produced about methods and about tools can often be falsified. The correctness of algorithms for the recognition of literary devices can be assessed, for instance, by comparing the expected results to the actual results, and, more precisely, by considering the level of precision and the level of recall. If it is accepted that the utility of text analysis tools can be assessed unequivocally, the field of literary informatics differs in an epistemological sense from the field of literary criticism. In this thesis, it was accepted that literary criticism does not aim to address questions via a single conclusive answer. Its primary aim is to continue a discussion about literary works, and new interpretations do not necessarily aim to invalidate previous interpretations. Chapters 6 and 8 of this thesis contain detailed discussions of the methodology that was followed, and this information was included in an attempt to contribute progressively to a knowledge about the nature and the value of algorithmic criticism.

10.2.2. Different scholarly output

The observations that practical work is generally based on theoretical assumptions, and that praxis is necessary to produce new theoretical insights, do not necessarily lead to the conclusion, however, that the non-textual resources that result from practical work can also function independently as a resource which can disseminate these theoretical insights. The act of modelling a physical object or a heuristic activity can in itself produce knowledge about the object or the activity being represented, but it is unclear if granting access to the software tool in which the model is implemented, or to a visualisation in which data is presented, simultaneously grants access to this knowledge. The humanities, like any other discipline, have developed standards for the ways in which knowledge may be communicated. The outcomes of enquiries are traditionally expounded in the form of discursive writing, and numerous authors have stressed that this is also the most effective channel.

This dominance of textual resources is increasingly being undercut within the digital humanities. Schnapp et al. argue that the digital humanities consists of “an array of convergent practices that explore a universe in which print is no longer the exclusive or the normative medium in which knowledge is produced and/or disseminated”.[608] Rockwell and Ramsay likewise draw attention to the fact that software tools can be viewed as resources which can independently proclaim a theory. In an attempt to establish “a materialist epistemology sufficient to the task of defending building as a distinct form of scholarly endeavour”,[609] the authors

[608] Jeffrey Schnapp, Peter Lunenfeld & Todd Pressner, *The Digital Humanities Manifesto 2.0*, p. 2.

[609] Stephen Ramsay & Geoffrey Rockwell, “Developing Things: Notes Towards an Epistemology of Building in the Digital Humanities”, in: Matthew K. Gold (ed.), *Debates in the Digital Humanities*, Minneapolis: University of Minnesota Press 2012, p. 77.

argue that software tools are "hermeneutical instruments through which we can interpret other phenomena". Like conceptual theories, digital tools enable scholars to deal with complexity by offering principles or guidelines to impose order on unorganised observations and to expose specific patterns or general qualities. Because of this quality, "text analysis and visualization tools are theories in the very highest tradition of what it is to theorize in the humanities".[610] Rockwell and Ramsay also argue that software applications can convey scholarly knowledge, as they can communicate specific ideas about the validity or the utility of innovative ways of presenting content. The digital humanities are centrally concerned with the development of new possibilities for engaging with the human record. While such new vistas may be described in words, the statement clearly gains rhetorical force when the ideas are actually embodied by a working application. The authors conclude that tools can prove a concept and that they can posit a thesis independently.

If software applications are to be recognised as genuine acts of scholarship, this demands a possibility for peers to critically respond to the argument that is presented. Mark Sample explains that "a creative or intellectual act becomes scholarship when it is public and circulates in a community of peers that evaluates and builds upon it".[611] Galey and Ruecker argue along similar lines that a digital artefact can be conceptualised as a scholarly object if it advances an argument, and, additionally, if this argument can be interpreted independently from any textual resources in which the resource is described. The authors' central proposition is that scholarly tools, like textual publications, can be subjected to peer review. Galey and Ruecker have developed a checklist which peers can use during the evaluation of digital tools. Amongst other criteria, it is stated that software tools ought to reify arguments which are "contestable, defensible, and substantive", it ought to have "a recognizable position in the context of similar work" and it should address possible objections.[612]

A theory that is expressed in code differs in a number of important ways, however, from a theory that is communicated in a discursive text. One crucial complication is that tools do not explicitly state their argument. The aims of the tools and the intentions of the developers mostly need to be decoded via a critical examination of the tool. The functionalities which are offered can often be gauged through actual usage, but to reconstruct the logic that is implemented, it is often necessary to have access to the source code. This code may be viewed as the modality in which the developer's insights are expressed. This communication via code limits the reading audience to readers who have a degree of proficiency in the

[610] Stephen Ramsay & Geoffrey Rockwell, "Developing Things: Notes Towards an Epistemology of Building in the Digital Humanities", p. 79.

[611] Mark Sample, "When Does Service Become Scholarship?", <http://www.samplereality.com/2013/02/08/when-does-service-become-scholarship/>., n.pag.

[612] A. Galey & S. Ruecker, "How a Prototype Argues", in: *Literary and Linguistic Computing*, 25:4 (27 October 2010), p. 414.

programming language that was used. A more serious difficulty, however, is that, even to those who can read the code, the tool can only ever reflect the intellectual efforts that were put into its development in an incomplete manner. Ramsey and Rockwell note that digital artefacts are often “insufficiently open about their theoretical underpinnings”.[613] In a discursive text, it is generally possible to admit to specific shortcomings or limitations of a theory, or to describe initial avenues of thinking which later proved to be unsuccessful. For a full and balanced evaluation of the reasoning that was followed, failures are usually as valuable as successes. Like scholarly arguments, algorithms have mostly evolved through cycles of trials and refutations, but, when initial bugs and flaws are removed, other scholars only have access to the version in which a functionality has been implemented successfully. When programmers make use of version management software, such trials and refutations can potentially be reconstructed by carefully comparing the different historical versions of the code. In most cases, however, such a contrastive comparison can highlight the changes, but not the motivation behind these changes. Tools generally contain a conclusion only, and no arguments in support of this conclusion. As a result, it is often difficult for peers to understand the reasoning that was followed during the creation of the code. Furthermore, the code in itself generally lacks information about the success rate of the algorithm. It may be the case, for instance, that the tool functions properly only in a limited number of cases.

As code cannot convey the full genesis nor the full rationale of an argument, such aspects need to be communicated via other channels. Fabretti suggests that software ought to be defined broadly as “the totality of all computer programs as well as all the written texts related to computer programs”. This definition covers not only the user interface and the underlying source code, but also the technical documentation and “the whole of technical literature related to computer programs, including methodological studies on how to design computer programs”.[614] The latter class of resources may be referred to as the epitext of software applications.[615] If such an expansive conceptualisation is accepted, the difficulties surrounding the legibility of software can be examined more effectively. Textual documentation about software is often necessary to outline particular misconceptions that may have existed prior to the full maturation of an algorithm. The applications in themselves usually lack a discussion of the assumptions that were held during the production process. They rarely convey a critical evaluation of their own performance.

[613] Stephen Ramsay & Geoffrey Rockwell, “Developing Things: Notes Towards an Epistemology of Building in the Digital Humanities”, p. 80.

[614] F. Fabretti, “Have the Humanities Always Been Digital?”, in: David Berry (ed.), *Understanding Digital Humaniities*, Basingstoke: Palgrave Macmillan 2012, p. 165.

[615] Gerard Genette & Marie Maclean, “Introduction to the Paratext”.

Since the digital humanities, to a large degree, focus on the development of methodology, [616] publications which explain and motivate why particular decisions were taken, and why specific alternatives have not been pursued, can serve an important function. Texts about the accuracy of tools promote a degree of transparency which is necessary for the evaluation of their suitability. Within humanistic discourse, texts which discuss the nature of algorithms are commonly regarded as being of a lesser rank, nevertheless. Scholars whose focus is predominantly on the formation or the application of theoretical concepts may presume that detailed ruminations on technical details do not belong naturally within the humanities, and may assert that questions associated with the extraction of data ought to be addressed instead within fields such as computer science or information science. The development of a method is sometimes viewed as a purely banausic activity, needed primarily as preparation for more evaluative work. While technical documentation is often viewed as a by-product of practical work, it seems clear that progress in the field of algorithmic criticism depends crucially on shared knowledge about the suitability of methods. De Roure notes that it is pivotal to share information on the methods by which results are generated. Such workflows used to produce a result "provides our route to repeatability, reproducibility and reuse". When such workflows are shared, they can also be "discussed and reviewed, reused and repurposed". De Roure also stresses that formal descriptions of workflows "are in many senses a new form of scholarly publication".[617]

Algorithms for the analysis of literary texts are currently still under development, and, to ensure that such work can be done effectively, it is crucial for scholars to share their insights about the accuracy of digital tools in scholarly texts. Scholars should contribute actively to the development of tools, so that they are not demoted to the role of mere observers. Through practical work, humanities researchers can ensure that technology is genuinely supportive of their research questions. Willard MacCarthy notes that experimentation places scholars "not merely in a position of witnesses or guessers but in the role of makers for whom the emergent potentialities of the medium constitute essential information".[618]

[616] Tom Scheinfeldt, "Where's the Beef? Does Digital Humanities Have to Answer Questions?", in: Matthew H. Gold (ed.), *Debates in the Digital Humanities*, Minneapolis: University of Minnesota Press 2012, p. 125.

[617] To support the sharing of workfows, Goble and De Roure have built the myExperiment website (www.myexperiment.org), which is "a social network of people sharing reusable methods for processing research data, in various research communities from bioinformatics and chemistry to climate change and digital humanities". It provides methods for the analysis of data. See David De Roure, Carole Goble & Robert Stevens, "The Design and Realisation of the Virtual Research Environment for Social Sharing of Workflows", in: *Future Generation Computer Systems*, 25:5 (2009).

[618] Willard McCarty, "Introduction", in: Willard McCarty (ed.), *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*, Oxford: Open Book Publishers 2010, p. 4.

Literary informatics, in conclusion, lead to different types of scholarly output. On the one hand, it results in a range of non-textual resources which expound their claims, mostly about their methodological assumptions, in an implicit form. On the other hand, the field also results in different types of textual publications, which document and which scrutinise the nature and the genesis of practical applications. Chapter 6, 7 and 8 of this thesis may be viewed as examples of this latter class of scholarly texts.

10.2.3. Different form of discovery

Machine reading enables scholars to observe patterns within collections in their entirety. The abstract phenomena that are observable at the macro-level are particularly valuable for studies in the field of literary history, of which Moretti's experiments with distant reading form clear examples. Visual representations of data can enable scholars to effectively investigate the synchronic or diachronic developments in phenomena such as genres or literary productivity. As has been explained, however, algorithmic criticism employs machine reading in support of literary criticism, which is a line of research which centrally aims to expose the unique properties of a singular work of literature. Within the context of literary criticism, a bare form of number crunching, which fully distances itself from the contents of the texts, is generally insufficient. The main value of abstract representations of collections lies in the fact that they can inspire more focused forms of readings at the microlevel. Exceptional data values can mostly be explained only by examining actual fragment in the texts which have produced these values. Martin Mueller argues that, while algorithmic processing can be applied initially to expose rough patterns within the corpus as a whole, the eventual aim of these abstractions is to suggest fragments which can subsequently be examined in more detail. Mueller uses the term "scalable reading" to describe the possibility to explore aspects of the corpus at different levels.[619]

Scholars who make use of digital methods to address the predicaments that result from abundance relegate the responsibility of making selections to the machine. The aim of algorithm-based filtering is typically to discriminate texts with relevant characteristics, or to expose aspects of texts which can in turn lead to new ideas about these texts. Algorithmic criticism is based on a mediated reading[620] in which algorithms are used to filter a text corpus. While the selection of texts often takes place haphazardly in analogue criticism, computation enables scholars to search methodically for texts which deviate from specific norms and which appear to warrant further reflection. Computational methods can mitigate the influence of

[619] Martin Mueller, *Stanley Fish and the Digital Humanities*, 2012, Martin Mueller, "Stanley Fish and the Digital Humanities", 2012, <http://cscdc.northwestern.edu/blog/?p=332> (12 March 2013).

[620] Shawna Ross, "In Praise of Overstating the Case: A Review of Franco Moretti, Distant Reading (London: Verso, 2013)", in: *Digital Humanities Quarterly*, 008:1 (2014), n.pag.

existing canons and can highlight texts which are distinctive on the basis of statistic grounds.

Algorithmic filtering is particularly useful if it results in unexpected selections. A number of authors have argued, nevertheless, that the algorithms that are used to filter large data collections can also reinforce existing subjective prejudices. Responding to Negroponte's prediction, made in 1995, that newspaper companies would develop personalised newspapers which exclusively contain the articles that are of relevance to a particular reader,[621] Sunstein expresses the concern that such forms of personalisation may lead to "information cocoons" in which "we hear only what we choose and only what comforts and pleases us".[622] In the *The Filter Bubble*, Pariser alleges in a similar vein that the manner in which we consume information is increasingly being determined by intermediaries which filter and organise this information. A crucial aspect of the type of filtering that is applied by search engines and social media platforms such as Google, Facebook and Twitter, however, is that it largely takes place outside of the awareness of their users. Pariser argues that, when personalisation is based on previous queries, this reinforces and sustains existing behaviour. Filtering mechanisms hide information which is unfamiliar to us and "indoctrinat[e] us with our own ideas".[623] When scholars devise their own algorithms, there may likewise be the risk that these mechanism spin an "information cocoon" in which the list of results unchangeably reflects the author's own interests. The particular algorithms that are chosen can subconsciously reinforce the existing preconceptions and expectations of the scholar. Ramsay argues that text mining in general is based on the assumption that the correct path towards the relevant information can be calculated. Such a logical and rational approach towards information retrieval may primarily produce results which are in step with a particular line of thinking. As such, text mining may frustrate serendipitous discoveries. As an alternative, Ramsay proposes a form of engagement which he refers to as "screwmeneutics".[624] Rather than enabling users to find objects via the process of filtering, digital libraries ought to facilitate an unrestrained navigation through the corpus. Users may have general interests, but they may not know beforehand which type of documents can actually meet these broad information needs.

The case study that was conducted as part of this thesis has shown, nevertheless, that machine reading can still lead to results which can unexpectedly

[621] Nicholas Negroponte, *Being Digital* (New York: Knopf 1995), p. 153.

[622] Cass Sunstein, *Infotopia: How Many Minds Produce Knowledge* (Oxford: Oxford University Press 2006), p. 9.

[623] Eli Pariser, *The Filter Bubble : What the Internet Is Hiding from You* (New York: Penguin Press 2011).

[624] Stephen Ramsay, "The Hermeneutics of Screwing Around; or What You Do with a Million Books", in: Kevin Kee (ed.), *Pastplay: Teaching and Learning History with Technology*, Ann Arbor: University of Michigan Press 2014.

initiate the thinking process. Due to the broad variability and the intrinsic unpredictability of literary phenomena, the results that are produced using a rule-based approach have error margins. For scholars who develop algorithms, it is generally impossible to foresee the full implications of a particular algorithm, and the results of such rule-based searches are likely to include unexpected text fragments. Whereas human being often have a clear notion of what makes a text significant or interesting, algorithms can only compute. This stalwart focus on quantification, in combination with inherent imperfection of algorithms, may fortuitously lead to serendipitous discoveries.

10.2.4. Different types of arguments

Algorithmic criticism is based on an alternative form of reading in which literary texts are converted into qualitative or quantitative data. Such data are generally used to gauge the differences and the similarities between these texts. Via digital methods, scholars can describe aspects of texts which are imperceptible to scholars who concentrate solely on paper-based resources. Examples of such supplementary stylistic indicators include the type-token ratio of a text, the text's average number of syllables per word or the standard deviation in the use of perfect rhyme within an entire volume of poetry. Whereas human analyses tend to focus on relatively limited collections of texts and on a relatively small set of literary devices within these texts, machine reading is a wholistic or an embrasive form of engagement in which the exact same types of metrics can be produced about the occurrences of widely diverse textual aspects such as repeated words, rhyme, grammar, metre and figures of speech. For human critics, it is generally difficult to be attentive to all of these aspects simultaneously, especially if some of these phenomena occur very frequently. Computer-based stylometric analyses may reveal, for instance, that the early work of a poet makes a very different use of pronouns than the later work of this poet. Such a distinction can be interesting from a critical point of view, but it is mostly strenuous to see such differences without computation. At the same time, it is also difficult for human readers to notice the absence of specific phenomena. Digital methods can easily establish, for instance, that some poems make a very extensive use of alliteration, while other poems are completely devoid of this device. Such relevant distinctions can easily be overlooked in conventional criticism.

Critics of the algorithmic approach often insist that these new forms of analyses rarely lead to relevant new insights, and that these methods merely confirm what is known or suspected already.[625] In answering this criticism, it is important to emphasise, firstly, that algorithmic criticism does not develop new questions in itself.

[625] Adam Kirsch, "Technology Is Taking Over English Departments: The False Promise of the Digital Humanities".

It needs to be viewed as a new methodology within the overarching field of literary studies, and it is meant to serve the same scholarly objectives. The aim of algorithmic criticism, like that of conventional criticism, is to arrive at a better understanding of the various aspects of literary works, such as their meaning, the relationship between form and content, and their relationship to other works in the same genre or literary period. Computational methods can be used to address questions which have likewise been investigated via conventional close reading, albeit in different ways. In the case study that was conducted for this thesis, for instance, a number of analyses have focused on the differences between MacNeice's poetry of the 1950s and the poetry written before and after this phase. When computational methods are applied to replicate traditional research, the findings of algorithmic criticism may either corroborate or repudiate the earlier findings. When quantitative analyses confirm what is known already, the very fact that a particular observation is confirmed by a fastidious computer-based analysis clearly adds authority to the scholarly claim. Because of the general differences in the overall methodology, because of the absence of subjective preferences for particular texts, and because of a general lack of knowledge about the historical or social context in which texts have originated, digital methods invariably answer these existing questions in fundamentally dissimilar ways.

The opposite situation, in which the results of computational methods contradict existing convictions, can be equally productive. Hugh Craig explains that the results of statistical processing can be especially interesting if they are surprising. Paradoxically, he expects to be "reassured by seeing patterns already familiar from the way texts are usually discussed, yet also to be surprised so that they seem more than a restatement of the obvious".[626] Conflicts between the expected results and the actual results may prompt scholars to find explanations for this discrepancy, and such additional analyses often lead to new ideas about the texts. As is also stressed by Jockers, the results of digital methods should not be viewed as conclusive evidence.[627] Answers obtained via quantitative methods may be still be contested via qualitative arguments. Ramsay stresses similarly that statistical processing has no more "claim to truth" than traditional forms of analysis.[628] Computational analyses can establish new perspectives from which texts can be analysed, and they can beneficially challenge accepted views. They can supply a range of new and disparate arguments which literary scholars may adopt to support and to undergird their scholarly claims.

[626] Hugh Craig, "Stylistic Analysis and Authorship Studies".

[627] Matthew Jockers, "So What?", *Matthew L. Jockers*, <http://www.matthewjockers.net/2014/05/07/so-what/> (7 May 2014).

[628] Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism*, pp. 14–15.

10.3. Continuities

10.3.1. Subjectivity

While there are clear differences between close reading and machine reading, the case study that was conducted for this thesis also underscored an important similarity. Various authors have claimed that the digital humanities can initiate a transition towards a more scientific form of scholarship, and that the empirical and factual observations that can allegedly be generated by computers can serve as a corrective to the subjectivity and the idiosyncrasy that prevails in many humanities disciplines. Moretti has argued, for instance, that distant reading is a form of textual engagement which rests "solidly on facts".[629] Rieder and Rohle explain that the perception of objectivity of digital method derives from the positivist conviction that the use of instruments, and the resultant eradication of human bias, leads to "results with a higher epistemological status".[630] The claim that computational methods automatically replace the subjective response of the individual reader with an objective scientific rigour seems deceptive, nevertheless, as the process of data creation is often steered to a large extent by subjective views. Algorithms for the generation of data are essentially hypotheses which speculate on the manner in which specific textual phenomena may be recognised. Data about the frequencies of tokens, for instance, demand a prior conceptualisation of the term "word". Different applications implement different rules for treating hyphenated words or compound nouns. Small modifications of these hypotheses may lead to widely different data. Algorithms ought to be viewed as cultural phenomena, as they are constructed by human software engineers who consciously or unconsciously take decisions on the types of results that they aim to produce. The phenomena which are annotated, moreover, do not constitute inherent properties of these texts. They are properties which scholars, working within a particular critical tradition, ascribe to these texts. Flanders and Jockers note astutely that "tools bring the data into existence, not just into view".[631] The data that are produced by text mining algorithms do not necessarily have a higher degree of objectivity than annotations which are compiled manually.

Like the procedures for the creation of data, the ways in which data sets are analysed are frequently driven by idiosyncratic or project-specific preferences. Analytic procedures such as clustering and the calculation of correlations have largely been standardised, and, as a result of this, the use of these statistical operations is often associated with increased objectivity. Particular forms of statis-

[629] Franco Moretti, *Distant Reading*, p. 44.

[630] Bernhard Rieder & Theo Rohle, "Digital Methods: Five Challenges", p. 72.

[631] Matthew Jockers & Julia Flanders, *A Matter of Scale. Keynote Lecture from the Boston Area Days of Digital Humanities Conference. Northeastern University, Boston, MA. March 18, 2013*, p. 16.

tical processing often bear the marks of individual proclivities, however. In the context of authorship attribution studies, for instance, researchers frequently eliminate specific variables from the analysis, in order to produce more compelling results. Scholars who explore texts statistically typically explore texts via different methods, and may notice that certain analytic methods yield more befitting results than other methods. As analytic methods are usually informed by specific expectations of what scholars hope to find, the results of statistical analyses should not necessarily be treated as irrefutable and objective evidence.

10.3.2. Alternation between inductive and deductive reasoning

Computational methods can in theory be applied to a corpus without any prior knowledge of the contents of the texts, and without any expectation of what these methods ought to yield. Analyses do not necessarily need to buttress a concrete research question, and they can initially be applied solely to search for specific patterns within the data. The fact that digital methods can be applied without a theoretical basis has frequently incited fierce criticism. Stanley Fish, for instance, repudiates the digital humanities in a series of blog posts for licensing free experimentation, and for attenuating the relevance of central research questions. Fish surmises that, while initial theories and hypotheses serve as indispensable search lights in humanities research, the focus within the digital humanities on the creation of abstract patterns which cannot be perceived directly by human readers impedes the formation of initial hypotheses. Since it is impossible to know beforehand which patterns will be produced by computer applications, the research cannot begin “in a motivated — that is, interpretively directed — way”. According to Fish, unmotivated experiments bear the risk of exposing “a correlation between a formal feature the computer program just happened to uncover and a significance that has simply been declared, not argued for”.[632] The unmotivated forms of research which are rejected by Fish are exemplified by the studies which are described in Chris Anderson’s essay “The End of Theory”. Anderson depicts explorations in which researchers randomly apply statistical procedures to big data sets, to find out only afterwards which hypotheses their results may support.[633]

The form of research that Fish advocates is essentially deductive in nature. Deduction departs from a central hypothesis, and aims to find data in support of the suggested proposition. Induction, by contrast, starts with the collection of observations, and aims to extract general principles or explanatory theories from these data. It is specious, however, to suggest that traditional research is exclusively deductive and that research driven by digital methods is exclusively

[632] Stanley Fish, “Mind Your P’s and B's: The Digital Humanities and Interpretation”, *New York Times*, 2012 <http://opinionator.blogs.nytimes.com/2012/01/23/mind-your-ps-and-bs-the-digital-humanities-and-interpretation/?_r=0>, n.pag.

[633] Chris Anderson, “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete”.

inductive. Kell and Oliver stress that deduction and induction ought to be viewed as complementary and equally valuable methods for producing knowledge, and note that most disciplines make use of a combination of inductive and deductive methods.[634] This is clearly the case for literary research as well, as a literary scholar rarely begins to read a new book with a fully developed hypothesis. In a first exploratory reading, critics usually search for remarkable passages or for specific reoccurring features. This initial examination may be viewed as an example of induction. On the basis of the qualities that were observed during the first reading, the scholar may develop a theory about this work, which can then be investigated in a deductive mode during subsequent, and more focused, encounters with the text.

Literary informatics research can likewise follow both inductive and deductive approaches. As was demonstrated during the case study that was conducted for this thesis, digital methods do not necessarily need to be motivated by an antecedent conjecture, and theoretical explanations can follow the formation of patterns. During initial explorations, scholars can cast their nets very widely and apply many different analytic techniques in a seemingly random manner, in order to generate patterns which can spark novel ideas. It is important to stress that such exploratory experiments seldom take place in a critical vacuum. In general, analyses can be productive only if they depart from a prior suspicion about relevant correlations and about meaningful patterns. The very design of experiments is generally based on a concrete research interest and on an initial curiosity. Researchers, importantly, need to take decisions on the texts that need to be compared, or on the variables that need to be correlated. A degree of knowledge about the general nature and the context of the corpus is clearly necessary to understand what makes patterns interesting in the first place. Dan Dixon shrewdly observes that this particular way of producing knowledge forms an adequate example of the general process which C.S. Peirce's referred to as abduction. Peirce proposed the term to formalise the hunches, suspicions and hypotheses that can be generated via the recognition of patterns and correlations in data collections. Kell and Oliver view abduction as a specific form of induction. Abduction is a random or haphazard process which is not based on logic, on formal rules or on a prior hypothesis.[635] As such, abductive reasoning curtails the preponderance of traditional research interests, and can function as a catalyst for neoteric ideas unskewed by conventional conceptions.

[634] Douglas B. Kell & Stephen G. Oliver, "Here Is the Evidence, Now What Is the Hypothesis? The Complementary Roles of Inductive and Hypothesis-Driven Science in the Post-Genomic Era", in: *BioEssays*, 26:1 (2004). According to Kell and Oliver, mathematics forms the only exception, as this area of research is exclusively inductive.

[635] Dan Dixon, "Analysis Tool or Research Methodology? Is There an Epistemology for Patterns?", in: David Berry (ed.), *Understanding Digital Humaniities*, Basingstoke: Palgrave Macmillan 2012, pp. 201–202.

By contrast, scholars can also design experiments specifically for the purpose of trying to corroborate or to refute pre-defined hypotheses. Contrary to what is claimed by Fish, the fact that scholars cannot know beforehand which patterns will be produced does not mean that they cannot perform experiments in support of a specific theory. Because algorithmic analyses can produce highly unexpected results, they often encourage a critical reflection on these hypotheses. It is often the case that experiments which flow naturally from a central hypothesis do not yield any meaningful results. The absence of such results can inspire scholars either to experiment with alternative analytic procedures, or to revise the initial theory. Statistical processing may expose qualities of the corpus which previously eluded the frame of reference of scholars, and such serendipitous findings can often lead to new experiments, conducted in an inductive fashion. In turn, such altered insights may spawn entirely new hypotheses. Like conventional criticism, algorithmic criticism frequently consists of an alternation of inductive and deductive approaches.

10.3.3. Critical reflection on the methodology

Like scholars using conventional methods, researchers who base their analyses on computation carry the responsibility to evaluate whether or not the methods they use can genuinely serve the central scholarly objectives of literary studies. Matching methods to specific pre-determined goals can be challenging, as technologies often set specific demands. Adherents of the technological determinism theory would assume that users of technology are woefully incapable of shaping the nature of their tools. In this extreme situation, the research agenda would be ruled entirely by what seems possible from a technical point of view, rather than by what is desirable from the perspective of literary criticism. Research which is principally driven by what can be studied rather than what should be studied is compared by Martin Mueller to “the old joke about the drunk who is looking for his lost car key under a lamp post because that is where the light is”.[636] The historian of science Thomas Hughes takes a stance which is less extreme, as he argues that users of technological systems often have a range of options concerning the way in which these technologies are applied. He emphases, nevertheless, that technological systems can still “acquire momentum”[637] at some point in their development. Under such conditions, opposing the dictates of technology is more exacting. Jacques Ellul argues along similar lines that technologies tend to be organised according to the “one best way”. Once it has been proven that a particular method ensures the maximal efficiciency, it becomes more difficult to alter

[636] Martin Mueller, “Digital Shakespeare, or towards a Literary Informatics”, p. 295.

[637] Thomas Hughes, “The Evolution of Large Technological Systems”, p. 76.

the course of technological development. Ellul described this tendency of technology to become self-directing using the phrase "automatism of technical choice".[638]

One of the central challenges within literary informatics is to move against the automatism that inheres in technological development. The digital technologies that are adopted by literary scholars are often accompanied by specific obligations or requirements which can have important consequences for the ways in which scholarly aims are realised. Using the TEI format, for example, demands a prior acceptance of the OHCO theory, and the use of RDF crucially demands what Stefan Gradmann refers to as "thinking in the graph".[639] Standardised text analysis tools similarly have the tendency to encourage particular types of research. They are typically based on textual aspects which can be detected with a degree of reliability, such as words, sentences or parts of speech. Because of this emphasis on formal textual aspects, text analysis tools often nudge scholars into the direction of stylometrics or authorship attribution research. They simultaneously discourage other forms of criticism, however, by not supplying any appropriate support. Tools invariably lack an out-of-the box support for performing feminist, Marxist, biographical or post-colonial readings of texts, for instance. Such forms of criticism may conceivably be boosted by creating a lexicon of terms with a Marxist connotation, or by building classifiers which can identify texts with a feminist slant, on the basis, for instance, of Naieve Bayes. Tools are crucially based on methodological and epistemological assumptions, and most of the existing text analysis tools implicitly assert the irrelevance of the critical approaches that are not supported. Scholars who identify such lacunae in the toolset ought to signal these shortcomings, and, if possible, they should carry out projects in which such deficiencies can be addressed. Without such a critical and practical engagement, the field will cease to evolve, causing a risk that particular approaches will be cemented as the disciplinary standard.

Algorithmic criticism demands programming skills and a proficiency in statistics, and such new competences are likely to have an impact on the manner in which scholars operationalise research questions. This development may potentially produce a number of undesirable effects. Wilkens fears that scholars who are frequently exposed to graphic renditions of data sets about text collections may partly lose their proficiency in traditional close reading.[640] Digital methods may stimulate scholars to analyse literary works predominantly in a mathematical manner, and to address questions of literary criticism in a facile manner by reducing these to differences and similarities which can be calculated. Since text mining necessarily focuses on textual aspects which can be detected algo-

[638] Jacques Ellul, *The Technological Society*, p. 79.

[639] Stephann Gradman, 'The Web & Digital Humanities: What about Semantics?', in *WW2012*, (Lyon), slide 2.

[640] Matthew Wilkens, "Canons, Close Reading, and the Evolution of Method", in: Matthew Gold (ed.), *Debates in the Digital Humanities*, Minneapolis: University of Minnesota Press 2012, p. 256.

rithmically, aspects which are difficult to quantify, such as the tenor of a metaphor, instances of ironical language or the connotations of words, may increasingly escape the scholar's radar.

An important risk that inheres in the adoption of digital methods is also that it can limit the scholarly focus to aspects which can be observed objectively and to claims which can be derived logically from these empirical observations. An adamant belief in the objectivity and the rationality of computation may undesirably lead to a restoration of the nineteenth century positivist belief that empirical and objective observations form that sole basis for reliable and authoritative knowledge. The attempt to present literary informatics as an approach which can unproblematically unearth the facts of a text is acutely out of step with current humanistic practices. Tymoczko stresses that the positivist aim of amassing facts "does not suffice in a post-positivist, globalizing world and will doom any field that adheres to such principles".[641] Within the humanities, the ideal of objective knowledge has largely been superseded by the insights that knowledge is perspectival and that human language can be arbitrary and ambiguous. Humanistic research focuses strongly "on multiplicity and ambiguity, on heterogeneity and difference".[642] Instead of merely concentrating on the rational aspects, this thesis has also stressed the subjective nature of algorithms, the methodological bias of text analysis tools and the continued need for human explication.

To ensure that computational methods can genuinely be of relevance to literary research, the functionalities that are offered by text analysis tools, and the methodology of literary informatics in general, must be scrutinised diligently and critically. The question of whether the outcomes of digital exertions are useful or meaningful can be evaluated, crucially, by connecting these to the central epistemological orientations of literary studies. The discipline is certainly not concerned solely with descriptive observations about texts. It also aims to interpret literary works, and to uncover the various layers of meaning that may exist within texts. David Levy usefully explains, more broadly, that there ought to be room for two distinct classes of activities within humanities research. Ratio refers to "the power of discursive thought, of searching and researching, abstracting, refining and concluding" while "intellectus refers to the ability of 'simply looking' to which the truth presents itself as a landscape presents itself to the eye".[643] Computers can partly automate rational tasks such as searching and filtering, and they can help scholars to make systematic descriptive analyses of texts. The critical process must not stop at making these observations, however. The patterns and the properties

[641] M. Tymoczko, "Will the Traditional Humanities Survive in the 21st Century?", in: *Organization*, 8:2 (1 May 2001), p. 290.

[642] Ibid.

[643] David Levy, "No Time to Think: Reflections on Information Technology and Contemplative Scholarship", in: *Ethics and Information Technology*, 9:9 (2007), p. 73. Levy's terminology is derived, in turn, from Joseph Pieper's book *Leisure, the Basis of Culture*.

that can be detected by computers ought to form the building blocks for the eventual interpretation and evaluation. To avoid a barren ossification of the field, algorithmic criticism should not focus exclusively on the rational aspects of the methodology. It ought to manifest itself clearly as a distinctly humanistic discipline, driven by the imperative to interpret, to explain and to criticise.

The validity of an interpretation cannot be computed, however, and activities such as reflection, interpretation or synthesis almost inevitably remain quintessentially human. The aim of literary informatics is not to make the human researcher redundant. By contrast, its critical limitations underscore the continued need for scholars who can perform higher criticism. This thesis has emphatically presented text mining not as an alternative but as an addendum to traditional scholarship.[644] It provides a supplementary range of methods which can enhance and enrich the existing discipline with new types of insights. In all cases, human critics continue to bear the responsibility for evaluating the relevance of the information that is extracted by digital research instruments. Tools can be used to generate hypotheses, but they cannot be used to prove them. Statistical analyses can provide the premises of an argument, but they cannot independently reach a conclusion from these premises.

Large collections of machine-readable texts, combined with the continuous advances in text technology, often arouse great expectations about new types of information and new types of insights, inaccessible to previous generations of scholars. This thesis has studied a number of ways in which the sundry possibilities that are offered by quantification and by algorithmic analyses may meaningfully be harnessed. As noted, discussions of the impact of computation tend to be highly positivist, and frequently highlight the widening of the scope and the acceleration of academic discoveries. In the spirit of such optimism, it may be stated that algorithmic criticism can veritably expand the scope and the diversity of literary research, by methodically exposing the structural and formal features of texts, and by facilitating studies that span different genres, different periods and different nationalities. At the same time, it seems clear that the digital medium also implies clear challenges and important restrictions. Capturing information requires "the discipline of expressing oneself within the limitations of computability".[645] Furthermore, the perfunctory and ratiocinative manner in which data are analysed appears to be in a stark opposition to other hermeneutic principles which are often valued in the field of literary research, such as empathy, intuition and serendipity. In stressing both the affordances and the limitations of literary informatics, this

[644] Matthew Jockers concurs that, although his work concentrates on analyses of large text corpora, analyses at the micro-level remain necessary for a large number of tasks. He uses the concept of "close mining" as an analogy. Excavating machines can be used to clear the pathway towards the places where diamonds can be found, by a human digger with a handpick is needed to actually reveal these gems. See Matthew Jockers, *Macroanalysis : Digital Methods and Literary History*, p. 171.

[645] John Unsworth, "Knowledge Representation in the Humanities", (1993).

thesis aimed to avoid both an undue positivism and a luddite techno-scepticism. Because of its crucial limitations, it is improbable that conventional close reading can ever be fully supplanted by machine reading. Because of the simultaneous affordances, however, literary informatics ought to be welcomed as a valuable additional method for studying the intricate effects that can be produced by literary works.

In "The Heresy of Paraphrase",[646] Cleanth Brooks stresses that, because well-written poems typically have a unique structure, consisting of meticulously balanced applications of literary techniques, any concise rendition of the text's meaning in an alternative phrasing is inevitably reductive. If the attempt to paraphrase a poem into plain prose is viewed as heretic, the aim to represent a literary work as a number, which is a central activity in algorithmic criticism, would likely be considered even less commendable by the New Critics. Such conversions into numerical data are generally needed, nonetheless, to allow for equitable comparisons of works of literature. All computer-based analyses must be preceded by a careful consideration of both the aspects which are quantified and the ways in which these aspects are quantified. Machine reading and data visualisation inherently imply abstraction and simplification, but such reductive methods are mostly applied for condonable reasons. The numbers which are generated do not form goals in themselves, as the ultimate objective of the various metrics is to reveal novel types of aspects and to spark fresh and startling ideas about the texts which are rendered numerically. Within texts which have already been examined closely and seemingly exhaustively, computational analyses may still discover characteristics which were previously unseen. Algorithmic criticism ultimately seeks to apply the power of computation to invigorate human interpretation, and to explore what can be gained from the heresy of quantification.

[646] "The Heresey of Paraphrase" is the final chapter in Cleanth Brooks, *The Well Wrought Urn: Studies in the Structure of Poetry*.

Bibliography

Works cited

A. Abdul-Rahman et al., "Rule-based Visual Mappings – with a Case Study on Poetry Visualization", in: B. Preim, P. Rheingans, & H. Theisel (eds.), *Eurographics Conference on Visualization (EuroVis)*, (2013), <http://dx.doi.org/10.1111/cgf.12125>.

Meyer Howard Abrams, *A Glossary of Literary Terms*, (Fort Worth: Harcourt Brace Jovanovich College Publishers 1993).

R. L. Ackoff, "From Data to Wisdom", in: *Journal of Applies Systems Analysis*, 16 (1989), pp. 3–9.

Sarah Allison et al., *Quantitative Formalism: An Experiment*, (Stanford: Stanford Literary Lab 2011), <http://litlab.stanford.edu/LiteraryLabPamphlet1.pdf>.

Richard Altick, *The Art of Literary Research*, (New York: Norton 1963).

Chris Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", in: *Wired Magazine*, 16:07 (2008), <http://www.wired.com/science/discoveries/magazine/16-07/pb_theory>.

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), pp. 3779–3796, <http://dx.doi.org/10.1098/rsta.2010.0156>.

Dawn Archer, Jonathan Culpeper & Paul Rayson, "Love – “a Familiar or a Devil”? An Exploration of Key Domains in Shakespeare’s Comedies and Tragedies", in: Dawn Archer (ed.), *What’s in a Word-List?: Investigating Word Frequency and Keyword Extraction*, (Farnham: Ashgate 2009), pp. 137–158.

Chris Baldick, *The Oxford Dictionary of Literary Terms*, (Oxford: 2009).

Roland Barthes, *Image, Music, Text*, (New York: Hill and Wang 1977).

Caroline Basset, "Canonicalism and the Computational Turn", in: David Berry (ed.), *Understanding Digital Humaniities*, (Basingstoke: Palgrave Macmillan 2012).

Marcia J. Bates, "Information and Knowledge: An Evolutionary Framework for Information Science", in: *Information Research*, 10:4 (2005), <http://www.informationr.net/ir/10-4/paper239.html>.

Jean Baudrillard, *Simulacra and Simulation*, (Ann Arbor: University of Michigan

Press 1994).

Jean Bauer, "Who You Calling Untheoretical?", in: *Journal of Digital Humanities*, 1:1 (2011), <http://journalofdigitalhumanities.org/1-1/who-you-calling-untheoretical-by-jean-bauer/>.

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).

Gordon Bell, Tony Hey & Alex Szalay, "Beyond the Data Deluge", in: *Science*, 323:5919 (6 March 2009), pp. 1297–8.

Tim Berners-Lee, James Hendler & Ora Lasilla, "The Semantic Web: A New Form of Web Content That Is Meaningful to Computers Will Unleash a Revolution of New Possibilities", in: *Scientific American*, (2001).

Tim Berners-Lee, "Foreword", in: Dieter Fensel (ed.), *Spinning the Semantic Web*, (Cambridge: MIT Press 2005).

Phillip William Berrie, *Just In Time Markup for Electronic Editions*, <http://www.unsw.adfa.edu.au/external_data_share/hass/ASEC/pdf/jitm/Wollongong200004PWB.pdf>.

David Berry, *Critical Theory and the Digital*, (London: Bloomsbury Publishing, 2014).

David M. Berry, "Introduction", in: *Understanding Digital Humanities*, (New York: Palgrave Macmillan 2012).

Jacques Bertin, *Semiology of Graphics*, (Madison: University of Wisconsin Press 1983).

Roger Bilisoly, *Practical Text Mining with Perl*, (Hoboken, N.J.: Wiley 2008).

Tobias Blanke & Mark Hedges, "Scholarly Primitives: Building Institutional Infrastructure for Humanities E-Science", in: *Future Generation Computer Systems*, 29:2 (February 2013), pp. 654–661, <http://dx.doi.org/10.1016/j.future.2011.06.006>.

David M. Blei, Andrew Y. Ng & Michael I. Jordan, "Latent Dirichlet Allocation", in: *The Journal of Machine Learning Research*, 3 (1 March 2003), pp. 993–1022, <http://dl.acm.org/citation.cfm?id=944919.944937>.

Jay Bolter, *Writing Space: The Computer, Hypertext, and the History of Writing*, (Hillsdale N.J.: L. Erlbaum Associates 1991).

Christine Borgman, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*, (Cambridge: MIT Press 2007).

Christine L. Borgman, "The Digital Future Is Now: A Call to Action for the Humanities", in: *Digital Humanities Quarterly*, 003:4 (2010), <http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html>.

J. Bradley, "Thinking about Interpretation: Pliny and Scholarship in the Humanities", in: *Literary and Linguistic Computing*, 23:3 (5 September 2008), pp. 263–279, <http://dx.doi.org/10.1093/llc/fqn021>.

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

Eric Brill, "A simple rule-based part of speech tagger", in: *Proceedings of the third conference on Applied natural language processing*, (Morristown, NJ, USA: Association for Computational Linguistics, 1992), p. 152.

Laurel Brinton, *The Structure of Modern English a Linguistic Introduction*, (Philadelphia: John Benjamins 2000).

Cleanth Brooks, *The Well Wrought Urn: Studies in the Structure of Poetry*, (London: Dennis Dobson 1968).

Cleanth Brooks, *Understanding Poetry*, (New York: Holt Rinehart and Winston 1960).

Susan Brown et al., "Reading Orlando with the Mandala Browser: A Case Study in Algorithmic Criticism via Experimental Visualization", in: *Digital Studies / Le champ numérique*, 2:1, <http://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/191/237>.

Ian Buchanan, *A Dictionary of Critical Theory*, (Oxford: Oxford University Press 2010).

Eric Bulson, "Ulysses by Numbers", in: *Representations*, 127 (2014).

J. Burrows, "Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship", in: *Literary and Linguistic Computing*, 17:3 (1 September 2002), pp. 267–287, <http://dx.doi.org/10.1093/llc/17.3.267>.

John Burrows, "Never Say Always Again: Reflections on the Numbers Game", in: Willard McCarty (ed.), *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*, (Cambridge: Open Book Publishers 2010).

John Burrows, "Textual Analysis", in: Susan Schreibman, Ray Siemens, & John Unsworth (eds.), *A Companion to Digital Humanities*, (Oxford: Blackwell 2002).

Roberto Busa, "The Annals of Humanities Computing: The Index Thomisticus", in:

Computers and the Humanities, 14 (1980), pp. 83–90.

Vannevar Bush, "As We May Think", in: *The Atlantic*, (1945), <http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/>.

Dino Buzzetti, "Digital Representation and the Text Model", in: *New Literary History*, 33:1 (2002), pp. 61–88.

Edward Carney, *A Survey of English Spelling*, (Routledge 1994).

Annamaria Carusi & Torsten Reimer, *VRE Collaborative Landscape Study*, (London: 2010), <http://www.jisc.ac.uk/media/documents/publications/vrelandscapereport.pdf>.

Manish Chaturvedi et al., "Myopia: A Visualization Tool in Support of Close Reading", (2012).

Min Chen & Luciano Floridi, "An Analysis of Information Visualisation", in: *Synthese*, 190:16 (26 September 2012), pp. 3421–3438.

Clara M. Chu, "Literary Critics at Work and Their Information Needs: A Research-Phases Model", in: *Library & Information Science Research*, 21:2 (January 1999), pp. 247–273.

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

Timothy Clark, "Interpretation: Hermeneutics", in: Patricia Waugh (ed.), *Literary Theory and Criticism: An Oxford Guide*, (Oxford University Press 2006).

T. E. Clement, "“A Thing Not Beginning and Not Ending”: Using Digital Tools to Distant-Read Gertrude Stein’s The Making of Americans", in: *Literary and Linguistic Computing*, 23:3 (5 September 2008), pp. 361–381.

N. Coffee et al., "The Tesserae Project: Intertextual Analysis of Latin Poetry", in: *Literary and Linguistic Computing*, 28:2 (20 July 2012), pp. 221–228, <http://dx.doi.org/10.1093/llc/fqs033>.

Neil Coffee et al., "Modelling the Interpretation of Literary Allusion with Machine Learning Techniques", in: *Digital Humanities 2013*, (Nebraska–Lincoln: 2013).

Margaret Cohen, "Narratology in the Archive of Literature", in: *Representations*, 108:1 (2009), pp. 51–75.

Beverley Collins & Inger Mees, *The Phonetics of English and Dutch*, (Leiden: Brill 1998).

Clare Connors, *Literary Theory*, (Oxford: Oneworld 2010).

James H. Coombs, Allen H. Renear & Steven J. DeRose, "Markup Systems and the Future of Scholarly Text Processing", in: *Communications of the ACM*, 30:11 (1987).

Karen Coyle, "FRBR, the Domain Model", in: *Library Technology Reports*, (2010), pp. 20–25.

Hugh Craig, "Stylistic Analysis and Authorship Studies", in: *A Companion to Digital Humanities*, (Oxford: Blackwell 2002).

Gregory Crane, "What Do You Do with a Million Books?", in: *D-Lib Magazine*, 12:3 (2006).

Gregory Crane & David Bamman, "The Logic and Discovery of Textual Allusion", in: *ACL Language Technology for Cultural Heritage*, (2008), <http://www.perseus.tufts.edu/~ababeu/latech2008.pdf>.

James Cummings, "The Text Encoding Initiative and the Study of Literature", in: *Blackwell Companion to Digital Literary Studies*, (Oxford: Blackwell 2007).

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

Robert I. Damper, "The Logic of Searle's Chinese Room Argument", in: *Minds and Machines*, 16:2 (18 October 2006), pp. 163–183.

Cathy N. Davidson, "What If Scholars in the Humanities Worked Together, in a Lab?", in: *The Chronicle of Higher Education*, 28 May 1999.

Cathy N. Davidson, "Humanities 2.0: Promise, Perils, Predictions", in: Matthew Gold (ed.), *Debates in the Digital Humanities*, (Minneapolis: University of Minnesota Press 2012).

R. Davis, H. Shrobe & P. Szolovits, "What Is a Knowledge Representation?", in: *AI Magazine*, 14:1 (1993), pp. 17–33.

Stanislas Dehaene, *Reading in the Brain: The Science and Evolution of a Human Invention*, (New York: Viking 2009).

Steven J. DeRose et al., "What Is Text, Really?", in: *Journal of Computing in Higher Education*, 1:2 (1990), pp. 2–26.

Dimitris A. Dervos & Anita Sundaram Coleman, "A Common Sense Approach to Defining Data, Information, and Metadata", in: Gerhard Budin, Christian Swertz, & Konstantin Mitgutsch (eds.), *Proceedings of the Ninth International ISKO Conference*, (Würzburg: Ergon-Verlag, 2006), pp. 51–58.

Cecily Devereux, "“A Kind of Dual Attentiveness”: Close Reading after the New Criticism", in: Miranda B Hickman & John D McIntyre (eds.), *Rereading the New Criticism*, (Columbus: Ohio State University Press 2012).

Joachim Diederich (ed.), *Rule Extraction from Support Vector Machines*, (Berlin: Springer 2008).

Wilhelm Dilthey, *Introduction to the Human Sciences*, (Princeton University Press 1989).

Dan Dixon, "Analysis Tool or Research Methodology? Is There an Epistemology for Patterns?", in: David Berry (ed.), *Understanding Digital Humaniities*, (Basingstoke: Palgrave Macmillan 2012).

Johanna Drucker, "Theory as Praxis: The Poetics of Electronic Textuality", in: *Digital Poetics*, (Tuscaloosa: University of Alabama Press 2002).

Johanna Drucker, "Humanities Approaches to Graphical Display", in: *Digital Humanities Quarterly*, 005:1 (2011), <http://digitalhumanities.org:8080/dhq/vol/5/1/000091/000091.html>.

Andrew DuBois, "Introduction", in: Frank Lentricchia & Andrew DuBois (eds.), *Close Reading: The Reader*, (Durham N.C.: Duke University Press 2003).

Terry Eagleton, *How to Read a Poem*, (Malden Mass.: Blackwell Pub. 2007).

Terry Eagleton, *Literary Theory: An Introduction*, (Minneapolis: University of Minnesota Press 1983).

M. Eder, "Does Size Matter? Authorship Attribution, Small Samples, Big Problem", in: *Literary and Linguistic Computing*, (14 November 2013), p. fqt066–, <http://dx.doi.org/10.1093/llc/fqt066>.

Paul Eggert, "The Book, the E-Text and the “Work-Site”", in: Marilyn Deegan & Kathryn Sutherland (eds.), *Text Editing, Print and the Digital World*, (Farnham: Ashgate 2009), pp. 63–83.

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).

Jacques Ellul, *The Technological Society*, (New York: Alfred A. Knopf 1973).

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

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).

Tom Eyers, "The Perils of the “Digital Humanities”: New Positivisms and the Fate of Literary Theory", in: *Postmodern Culture*, 23:2 (2013).

F. Fabretti, "Have the Humanities Always Been Digital?", in: David Berry (ed.), *Understanding Digital Humaniities*, (Basingstoke: Palgrave Macmillan 2012).

Lucien Febvre & Henri-Jean Martin, *The Coming of the Book: The Impact of Printing 1450-1800*, (London: NLB 1976).

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

Stephen Few, "Data Visualization for Human Perception", in: Mads Soegaard & Rikke Friis Dam (eds.), *The Encyclopedia of Human-Computer Interaction*, (Aarhus: The Interaction Design Foundation 2014).

Stanley Fish, "Mind Your P’s and B's: The Digital Humanities and Interpretation", in: *New York Times*, 2012 <http://opinionator.blogs.nytimes.com/2012/01/23/mind-your-ps-and-bs-the-digital-humanities-and-interpretation/?_r=0>.

Kathleen Fitzpatrick, "The Humanities, Done Digitally", in: *Debates in the Digital Humanities*, (University of Minnesota Press 2012).

Julia Flanders, "The Productive Unease of 21st-Century Digital Scholarship", in: Melissa Terras, Julianne Nyhan, & Edward Vanhoutte (eds.), *Defining Digital Humanities*, (Farnham: Ashgate 2013), pp. 205–219.

Richard S Forsyth, "Stylochronometry with Substrings, or: A Poet Young and Old", in: *Literary and Linguistic Computing*, 14:4 (1999), pp. 467–477, <http://dx.doi.org/10.1093/llc/14.4.467>.

Louise Francis, "Taming Text: An Introduction to Text Mining", in: *Casualty Actuarial Society Forum*, (2006), <http://www.casact.net/pubs/forum/06wforum/06w55.pdf>.

Martin Frické, "The Knowledge Pyramid: A Critique of the DIKW Hierarchy", in: *Journal of Information Science*, 35:2 (21 November 2008), pp. 131–142.

Ben Fry, *Visualizing Data*, (Cambridge: O’Reilly Media Inc. 2008).

Northrop Frye, *Anatomy of Criticism*, (Princeton, New Jersey: Princeton University Press 1957).

Hans-Georg Gadamer, *Truth and Method*, (New York: Seabury Press 1975).

A. Galey & S. Ruecker, "How a Prototype Argues", in: *Literary and Linguistic Computing*, 25:4 (27 October 2010), pp. 405–424, <http://dx.doi.org/10.1093/llc/fqq021>.

Jane Gallop, "The Historicization of Literary Studies and the Fate of Close Reading", in: *Profession*, (2007), pp. 181–186.

Jane Gallop, "The Ethics of Reading", in: *Journal of Curriculum Theorizing*, (2000).

Gerard Genette & Marie Maclean, "Introduction to the Paratext", in: *New Literary History*, 22:2 (2010), pp. 261–272.

Frederick W. Gibbs & Trevor J. Owens, "The Hermeneutics of Data and Historical Writing", in: Kristen Nawrotzki & Jack Dougherty (eds.), *Writing History in the Digital Age*, (MIT Press 2013).

Sigfried Giedion, *Mechanization Takes Command: A Contribution to Anonymous History.*, (Oxford University Press 1948).

P. Gooding, M. Terras & C. Warwick, "The Myth of the New: Mass Digitization, Distant Reading, and the Future of the Book", in: *Literary and Linguistic Computing*, 28:4 (13 August 2013), pp. 629–639, <http://dx.doi.org/10.1093/llc/fqt051>.

Stephann Gradman, "The Web & Digital Humanities: What about Semantics?", in: *WW2012*, (Lyon: 2012).

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

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).

Keith Grint, *The Machine at Work: Technology, Work, and Organization*, (Cambridge: Polity Press 1997).

Michael Groden, *The Johns Hopkins Guide to Literary Theory and Criticism*, (Baltimore: Johns Hopkins University Press 1994).

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).

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>.

R. R. K. Hartmann & Gregory James, *Dictionary of Lexicography*, (Routledge 2002).

Katherine Hayles, *How We Think: Digital Media and Contemporary Technogenesis*, (Chicago: The University of Chicago Press 2012).

Malcolm Hayward, "Analysis of a Corpus of Poetry by a Connectionist Model of Poetic Meter", in: *Poetics*, 24:1 (July 1996), pp. 1–11, <http://www.sciencedirect.com/science/article/pii/0304422X95000129>.

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).

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), p. S5, <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-S14-S5>.

J. Berenike Herrmann, Karina van Dalen-Oskam & Christof Schöch, "Revisiting Style, a Key Concept in Literary Studies", in: *Journal of Literary Theory*, 9:1 (2015), pp. 25–52, <http://www.jltonline.de/index.php/articles/article/view/757/1764>.

Ryan Heuser & Long Le-Khac, *A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method*, (Stanford Literary Lab 2012), <http://litlab.stanford.edu/LiteraryLabPamphlet4.pdf>.

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), pp. 809–824.

Miranda B. Hickman, "Introduction: Rereading the New Criticism", in: Miranda B. Hickman & John D. McIntyre (eds.), *Rereading the New Criticism*, (Columbus: Ohio State University Press 2012).

Susan Hockey, *Electronic Texts in the Humanities: Principles and Practice*, (Oxford; New York: Oxford University Press 2000).

D. L. Hoover, "Testing Burrows's Delta", in: *Literary and Linguistic Computing*, 19:4 (1 November 2004), pp. 453–475, <http://dx.doi.org/10.1093/llc/19.4.453>.

David Hoover, "Quantitative Analysis and Literary Studies", in: Susan Schreibman & Ray Siemens (eds.), *A Companion to Digital Literary Studies*, (Oxford: Blackwell 2008).

David Hoover, "Word Frequency, Statistical Stylistics and Authorship Attribution", in: Dawn Archer (ed.), *What's in a Word-List?: Investigating Word Frequency and Keyword Extraction*, (Farnham: Ashgate 2009), pp. 35–52.

David Hoy, *The Critical Circle: Literature, History, and Philosophical Hermeneutics*, (Berkeley: University of California Press 1978).

Thomas Hughes, "The Evolution of Large Technological Systems", in: Wiebe Bijker, Thomas Hughes, & Trevor Pinch (eds.), *The Social Construction of Technological Systems: New Dirextions in the Sociology and History of Technology*, (Cambridge: MIT Press 1987), pp. 51–82.

Claus Huitfeldt, "Multi-Dimensional Texts in a One-Dimensional Medium", in: *Computers and the Humanities*, 28 (1994), pp. 235–241.

Jessica Hullman & Nicholas Diakopoulos, "Visualization Rhetoric: Framing Effects in Narrative Visualization", in: *IEEE Transactions on Visualization and Computer Graphics*, (2011).

IFLA, *Functional Requirements for Bibliographic Records: Final Report*, (2009), <http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf>.

Roman Jakobson, "Closing Statement: Linguistics and Poetics", in: Thomas A Sebeok (ed.), *Style in Language*, Advances in Semiotics, (Cambridge, Mass: MIT Press 1960), pp. 350–377.

G. James et al., *An Introduction to Statistical Learning, with Applications in R*, (Springer 2013).

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

M. Jessop, "Digital Visualization as a Scholarly Activity", in: *Literary and Linguistic Computing*, 23:3 (2008), pp. 281–293, <http://dx.doi.org/10.1093/llc/fqn016>.

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

Matthew Jockers, "So What?", *Matthew L. Jockers*, 7 May 2014, <http://www.matthewjockers.net/2014/05/07/so-what/>.

Matthew Jockers, *Text Analysis with R for Students of Literature* (Berlin: Springer, 2014).

Matthew Jockers & Julia Flanders, *A Matter of Scale. Keynote Lecture from the Boston Area Days of Digital Humanities Conference. Northeastern University, Boston, MA. March 18, 2013*, <http://digitalcommons.unl.edu/englishfacpubs/106>.

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

Sarah Jones, "When Computers Read: Literary Analysis and Digital Technology", in: *Bulletin of the American Society of Information Science and Technology*, 38:4 (2012).

Daniel Jurafsky & James H. Martin, *Speech and Language Processing*, (Englewood Cliffs: Prentice Hall 2008).

Joel Kalvesmaki, "Canonical References in Electronic Texts: Rationale and Best Practices", in: *Digital Humanities Quarterly*, 008:2 (2014), <http://www.digitalhumanities.org/dhq/vol/8/2/000181/000181.html>.

Justine Kao & Dan Jurafsky, "A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry", in: *NAACL Workshop on Computational Linguistics for Literature*, (2012), pp. 8–7.

David Kaplan & D.M. Blei, "A Computational Approach to Style in American Poetry", in: *Seventh IEEE International Conference on Data Mining*, (2007), pp. 553–558.

David Maxwell Kaplan, *Computational Analysis and Visualized Comparison of Style in American Poetry*, (Princeton University 2006).

Douglas B. Kell & Stephen G. Oliver, "Here Is the Evidence, Now What Is the Hypothesis? The Complementary Roles of Inductive and Hypothesis-Driven Science in the Post-Genomic Era", in: *BioEssays*, 26:1 (2004), pp. 99–105.

Mills Kelly, "Making Digital Scholarship Count", in: Dan Cohen & Tom Scheinfeldt (eds.), *Hacking the Academy: A Book Crowdsourced in One Week*, (Michigan: MPublishing 2011).

Kathleen Kerr & Waqas Javen, "Visualization and Rhetoric: Key Concerns for Utilizing Big Data in Humanities Research", in: *IEEE International Conference on Big Data*, (2013).

John M. Kirk, "Word Frequency: Use or Misuse?", in: Dawn Archer (ed.), *What's in a Word-List?: Investigating Word Frequency and Keyword Extraction*, (Farnham: Ashgate 2009), pp. 17–34.

Adam Kirsch, "Technology Is Taking Over English Departments: The False Promise of the Digital Humanities", in: *New Republic*, :May 2 (2014), <https://newrepublic.com/article/117428/limits-digital-humanities-adam-kirsch>.

Matthew Kirschenbaum, "The Remaking of Reading: Data Mining and the Digital Humanities".

Rob Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*, (2014).

Mario Klarer, *An Introduction to Literary Studies*, (London: Routledge 1999).

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

<http://www.textgrid.de/fileadmin/berichte-2/report-4-3-1.pdf>.

Lee Lacy, *OWL: Representing Information Using the Web Ontology Language*, (Victoria BC Canada: Trafford 2005).

Douglas Laney, *3D Data Management: Controlling Data Volume, Velocity, and Variety*, (2001).

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

Frank Raymond Leavis, *The Great Tradition: George Eliot, Henry James, Joseph Conrad*, ([New York]: New York University Press 1963).

John Lennard, *The Poetry Handbook: A Guide to Reading Poetry for Pleasure and Practical Criticism*, (Oxford [u.a.]: Oxford Univ. Press 2005).

Frank Lentricchia, "Preface", in: Frank Lentricchia & Andrew DuBois (eds.), *Close Reading: The Reader*, (Durham: Duke University Press 2002).

David Levy, "No Time to Think: Reflections on Information Technology and Contemplative Scholarship", in: *Ethics and Information Technology*, 9:9 (2007).

Anthony Liew, "Understanding Data, Information, Knowledge And Their Inter-Relationshipso Title", in: *Journal of Knowledge Management Practice*, 8:2 (2007).

A. Liu, "The State of the Digital Humanities: A Report and a Critique", in: *Arts and Humanities in Higher Education*, 11:1-2 (1 December 2011), pp. 8–41, <http://dx.doi.org/10.1177/1474022211427364>.

Alan Liu, "Where Is Cultural Criticism in the Digital Humanities?", in: Matthew Gold (ed.), *Debates in the Digital Humanities*, (University of Minnesota Press 2012).

Liz Lyon, *Dealing with Data: Roles, Rights, Responsibilities and Relationships*, (Bath: 2007).

Paul Maas, *Textual Criticism*, (Oxford: Clarendon Press 1958).

Tim William Machan, "Late Middle English Texts and the Higher and Lower Criticisms", in: Tim William Machan (ed.), *Medieval Literature: Texts and Interpretation*, (Binghamton: Center for Medieval and Early Renaissance Studies 1991).

K. Mahowald, "A Naive Bayes Classifier for Shakespeare's Second-Person Pronoun", in: *Literary and Linguistic Computing*, 27:1 (10 November 2011), pp. 17–23.

Lev Manovich, "How to Compare One Million Images?", in: *Understanding Digital Humanities*, (New York: Palgrave Macmillan 2012).

Lev Manovich, *The Language of New Media*, (Cambridg, Mass.: The MIT Press 2002).

Lev Manovich, "What Is Visualization?", in: *Visual Studies*, (2011).

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

Margaret Masterman, "The Intellect's New Eye", in: *Freeing the Mind: Articles and Letters from The Times Literary Supplement during March-June 1962*, (London: Times Publishing Company Ltd. 1962), pp. 38–44.

Stephen Matterson, "The New Criticism", in: *Literary Theory and Criticism: An Oxford Guide*, (Oxford: Oxford University Press 2006).

Willard McCarty, *Humanities Computing*, (Basingstoke; New York: Palgrave Macmillan 2005).

Willard McCarty, "Modeling: A Study in Word and Meaning", in: Susan Schreibman, Ray Siemens, & John Unsworth (eds.), *A Companion to Digital Humanities*, (Blackwell).

Willard McCarty, "A Telescope for the Mind?", in: Matthew Gold (ed.), *Debates in the Digital Humanities*, (Minneapolis: University of Minnesota Press 2012).

Willard McCarty, "Introduction", in: Willard McCarty (ed.), *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*, (Oxford: Open Book Publishers 2010), pp. p. 1–11.

Neil McCaw, *How to Read Texts: A Student Guide to Critical Approaches and Skills*, (London: Continuum 2008).

Jerome McGann, *Radiant Textuality: Literature after the World Wide Web*, (New York: Palgrave Macmillan 2004).

Robert E. McGrath, *XML and Scientific File Formats*, (Urbana-Champaign: 2003).

Donald Francis McKenzie, *Bibliography and the Sociology of Texts*, (Cambridge: Cambridge University Press 1999).

Jan Christoph Meister et al., "Crowdsourcing meaning: a hands-on introduction to CLÉA, the Collaborative Literature Éxploration and Annotation Environment", in: *Digital Humanities 2012*, (Hamburg: 2012), <http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/crowdsourcing-meaning-a-

hands-on-introduction-to-clea-the-collaborative-literature-exploration-and-annotation-environment.html>.

Albert Meroño-Peñuela, "Semantic web for the humanities", in: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, (Springer, 2013).

Barend Mons & Jan Velterop, "Nano-Publication in the e-science era", (2009).

Franco Moretti, *Distant Reading*, (London: Verso 2013).

Franco Moretti, *Graphs, Maps, Trees: Abstract Models for a Literary History*, (Verso 2005).

Martin Mueller, "Stanley Fish and the Digital Humanities", 2012, <http://cscdc.northwestern.edu/blog/?p=332>.

Martin Mueller, "Digital Shakespeare, or towards a Literary Informatics", in: *Shakespeare*, 4:3 (September 2008), pp. 284–301, <http://www.tandfonline.com/doi/abs/10.1080/17450910802295179>.

Lewis Mumford, *Technics and Civilization*, (New York: Harcourt Brace and Co. 1934).

N. M. Luscombe, D. Greenbaum & M. Gerstein, "What Is Bioinformatics? A Proposed Definition and Overview of the Field", <http://archive.gersteinlab.org/papers/e-print/whatis-mim/text.pdf>.

Nicholas Negroponte, *Being Digital*, (New York: Knopf 1995).

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

Mary Oliver, *A Poetry Handbook*, (San Diego: Harcourt Brace & Co. 1994).

David Olson, *The World on Paper: The Conceptual and Cognitive Implications of Writing and Reading*, (Cambridge: Cambridge University Press 1994).

Walter Ong, *Orality and Literacy: The Technologizing of the Word*, (London: Routledge 2002).

Carole L. Palmer, Lauren C. Teffeau & Carrie M. Pirmann, *Scholarly Information Practices in the Online Environment: Themes from the Literature and Implications for Library Service Development.*, (2009).

Eli Pariser, *The Filter Bubble : What the Internet Is Hiding from You*, (New York: Penguin Press 2011).

Willie van Peer, "Where Do Literary Themes Come From?", in: Max Louwerse & Willie van Peer (eds.), *Thematics: Interdisciplinary Studies*, (Amsterdam/Philadelphia: John Benjamins 2002).

Majorie Perloff, *Differentials: Poetry, Poetics, Pedagogy*, (University of Alabama Press 2004).

Perry Willett, "Electronic Texts: Audiences and Purposes", in: Susan Schreibman, Ray Siemens, & John Unsworth (eds.), *Blackwell Companion to Digital Humanities*, (Oxford: Blackwell), p. 2002.

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).

Elena Pierazzo, *Digital Scholarly Editing: Theories, Models and Methods*, (Farnham: Ashgate 2015).

Adam Piette, "Contempory Poetry and Close Reading", in: Peter Robinson (ed.), *The Oxford Handbook of Contemporary British and Irish Poetry*, (Oxford: Oxford University Press 2013).

Plato, *Phaedrus*, (Cambridge: Harvard University Press 1953).

Michael Polanyi, *Personal Knowledge Towards a Post-Critical Philosophy.*, (Chicago: University of Chicago Press 1958).

Ezra Pound, *ABC of Reading*, (London: Faber and Faber 1991).

Jessica Pressman, *Digital Modernism: Making It New in New Media*, (Oxford: Oxford University Press 2014).

Vladimir Propp, *Morphology of the Folktale*, (Austin: University of Texas Press 1968).

Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism*, (Urbana: University of Illinois Press 2011).

Stephen Ramsay, "The Hermeneutics of Screwing Around; or What You Do with a Million Books", in: Kevin Kee (ed.), *Pastplay: Teaching and Learning History with Technology*, (Ann Arbor: University of Michigan Press 2014).

Stephen Ramsay & Geoffrey Rockwell, "Developing Things: Notes Towards an Epistemology of Building in the Digital Humanities", in: Matthew K. Gold (ed.), *Debates in the Digital Humanities*, (Minneapolis: University of Minnesota Press 2012), pp. 75–84.

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

John Crowe Ransom, "Criticism Inc.", in: *The Virginia Quarterly Review*, :Autumn (1937), pp. 586–602.

Allen Renear, David Dubin & C. M. Sperberg-McQueen, "Towards a semantics for XML markup", in: *Proceedings of the 2002 ACM symposium on Document engineering - DocEng '02*, (New York, New York, USA: ACM Press, 2002), p. 119, <http://dl.acm.org/citation.cfm?id=585058.585081>, 12 November 2015.

Allen Renear, David Durand & Elli Mylonas, "Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies", in: *Research in Humanities Computing*, (Oxford: Oxford University Press 1995).

Allen H. Renear, "Text Encoding", in: *A Blackwell Companion to Digital Humanities*, (Oxford: Blackwell 2002).

Lisa M. Rhody, "Topic Modeling and Figurative Language", in: *Journal of Digital Humanities*, 2:1 (2012), <http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/>.

Ivar Armstrong Richards, *Principles of Literary Criticism.*, (New York: Harcourt Brace 1961).

Ivar Armstrong Richards, *Practical Criticism : A Study of Literary Judgment*, (London: K. Paul Trench Trubner 1929).

Allen Riddell, "How to Read 22,198 Journal Articles: Studing the History of German Studies with Topic Models", in: Matt Erlin (ed.), *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, (Boydell & Brewer Ltd 2014).

Bernhard Rieder & Theo Rohle, "Digital Methods: Five Challenges", in: David Berry (ed.), *Understanding Digital Humaniities*, (Basingstoke: Palgrave Macmillan 2012).

Dirk Roorda & Charles van den Heuvel, "Annotation as a New Paradigm in Research Archiving".

Shawna Ross, "In Praise of Overstating the Case: A Review of Franco Moretti, Distant Reading (London: Verso, 2013)", in: *Digital Humanities Quarterly*, 008:1 (2014).

David De Roure, Carole Goble & Robert Stevens, "The Design and Realisation of the Virtual Research Environment for Social Sharing of Workflows", in: *Future Generation Computer Systems*, 25:5 (2009), pp. 561–567.

J. Rowley, "The Wisdom Hierarchy: Representations of the DIKW Hierarchy", in: *Journal of Information Science*, 33:2 (February 2007), pp. 163–180, <http://jis.sagepub.com/cgi/content/abstract/33/2/163>.

Stan Ruecker, "Rich Prospect Browsing Interfaces", in: *Affordances of Prospect for Academic Users of Interpretively Tagged Text Collections*, (University of

Alberta 2003).

Paul Saenger, *Space between Words : The Origins of Silent Reading*, (Stanford Calif.: Stanford University Press 1997).

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).

Mark Sample, "When Does Service Become Scholarship?", <http://www.samplereality.com/2013/02/08/when-does-service-become-scholarship/>.

T. Saracevic, "Information Science", in: M. J. Bates (ed.), *Encyclopedia of Library and Information Sciences*, 3rd editio, (New York: Taylor and Francis), pp. 2570–2585.

Ferdinand de Saussure, *Course in General Linguistics*, (London: Duckworth 1983).

Tom Scheinfeldt, "Why Digital Humanities is “Nice”", in: Matthew K. Gold (ed.), *Debates in the Digital Humanities*, (Minneapolis: University of Minnesota Press 2012).

Tom Scheinfeldt, "Where’s the Beef? Does Digital Humanities Have to Answer Questions?", in: Matthew H. Gold (ed.), *Debates in the Digital Humanities*, (Minneapolis: University of Minnesota Press 2012).

Jeffrey Schnapp, Peter Lunenfeld & Todd Pressner, *The Digital Humanities Manifesto 2.0*, (2009).

Chrostoph Schöch, "Big? Smart? Clean? Messy? Data in the Humanities", in: *Journal of Digital Humanities*, 2:3 (2013), <http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/>.

Robert Scholes, *Textual Power: Literary Theory and the Teaching of English*, (New Haven: Yale University Press 1985).

Susan Schreibman, Ray Siemens, & John Unsworth (eds.), *A Companion to Digital Humanities*, (Malden, MA: Blackwell 2004).

Susan Schreibman, Ray Siemens, & John Unsworth (eds.), *A Companion to Digital Literary Studies*, (Malden, MA: Blackwell 2007).

Erik Schultes & Mark Thompson, "Using Nanopublications to Incentivize the Semantic Exposure of Life Science Information".

David Schur, "An Introduction to Close Reading", (1998).

Cesare Segre, *Introduction to the Analysis of the Literary Text*, (Bloomington: Indiana University Press 1988).

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

Stéfan Sinclair, Stan Ruecker & Milena Radzikowska, "Information Visualization for Humanities Scholars", in: *Literary Studies in the Digital Age*, (Modern Language Association of America 2013).

Kate Singer, "Digital Close Reading: TEI for Teaching Poetic Vocabularies", in: *Journal of Interactive Technology and Pedagogy*, :3, <http://jitp.commons.gc.cuny.edu/digital-close-reading-tei-for-teaching-poetic-vocabularies/>.

Robin Skelton, "Celt and Classicist: The Versecraft of Louis MacNeice", in: Terence Brown & Alec Reid (eds.), *Time Was Away*, (Dublin: Dolmen Press 1974), pp. 43–53.

Philip Smallwood, "Criticism, Valuation, and Useful Purpose", in: *New Literary History*, 28:4 (1 November 1997), pp. 711–722, <http://muse.jhu.edu/journals/new_literary_history/v028/28.4smallwood.html>.

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).

Martha Nell Smith et al., "“Undiscovered Public Knowledge”: Mining for Patterns of Erotic Language in Emily Dickinson’s Correspondence with Susan Huntington (Gilbert) Dickinson", in: *Digital Humanities*, (2006), pp. 252–255.

John F. Sowa, *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, (Brooks / Cole 1999).

George Steiner, *Real Presences*, (University Of Chicago Press 1991).

Maureen Stone, "Information Visualization: Challenge for the Humanities", in: *Working Together or Apart : Promoting the next Generation of Digital Scholarship : Report of a Workshop Cosponsored by the Council on Library and Information Resources and the National Endowment for the Humanities*, (Washington D.C.: 2009), pp. 43–57.

Cass Sunstein, *Infotopia: How Many Minds Produce Knowledge*, (Oxford: Oxford University Press 2006).

Kathryn Sutherland, "Being Critical: Paper-Based Editing and the Digital Environment", in: Marilyn Deegan & Kathryn Sutherland (eds.), *Text Editing,* *Print and the Digital World*, (Farnham: Ashgate).

G. Thomas Tanselle, *Textual Criticism since Greg: A Chronicle, 1950-1985*, (Charlottesville: University Press of Virginia 1987).

Verena Theile, "New Formalism(s): A Prologue", in: Verena Theile & Linda Tredennick (eds.), *New Formalisms and Literary Theory*, (Basingstoke: Palgrave Macmillan 2013).

Alex Thompson, "Deconstruction", in: Patricia Waugh (ed.), *Literary Theory and Criticism*, (Oxford: Oxford University Press 2006).

Katie Trumpener, "Critical Response: I. Paratext and Genre System: A Response to Franco Moretti", in: *Critical Inquiry*, 36:1, pp. 159–71, <http://emc.english.ucsb.edu/emc-courses/Novel-Mediation-S2011/novel-mediation/Articles/Moretti.style.inc.Tumpener.reply.pdf>.

Edward Tufte, *The Visual Display of Quantitative Information*, (Cheshire: Graphics Press 1983).

Alan Turing, "Computing Machinery and Intelligence", in: Noah Wardrip-Fruin & Nick Montfort (eds.), *The New Media Reader*, (Cambridge Mass.: The MIT Press 2003).

Žiga Turk et al., *ICT Ontological Framework and Classification*, (Ljubljana: 2002).

M. Tymoczko, "Will the Traditional Humanities Survive in the 21st Century?", in: *Organization*, 8:2 (1 May 2001), pp. 285–297.

John Unsworth, "What Is Humanities Computing and What Is Not?", in: Melissa Terras, Julianne Nyhan, & Edward Vanhoutte (eds.), *Defining Digital Humanities: A Reader*, (2013).

John Unsworth, "Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?".

John Unsworth, "Knowledge Representation in the Humanities", (1993), <http://www3.isrl.illinois.edu/~unsworth/KR/KRinHC.html>.

John Unsworth & Martin Mueller, *The MONK Project Final Report*, (2009).

Edward Vanhoutte, "Every Reader His Own Bibliographer - an Absurdity?", in: Marylin Deegan & Kathryn Sutherland (eds.), *Text Editing, Print and the Digital World*, (Farnham: Ashgate 2009).

Harold Veeser, *The New Historicism*, (New York: Routledge 1989).

Kim H. Veltman, "Towards a Semantic Web for Culture", in: *Journal of digital information*, 4:4 (2006).

Martin Volk, Lenz Furrer & Rico Sennrich, "Strategies for Reducting and Correcting OCR Errors", in: Caroline Sporleder, Kalliope Zervanou, & Antal van den Bosch (eds.), *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, (Berlin: Springer 2011).

Patricia Waugh, "Value: Criticism, Canons and Evaluation", in: Patricia Waugh (ed.), *Literary Theory and Criticism: An Oxford Guide*, (Oxford: Oxford University Press 2006).

Adriaan van der Weel, "Pandora's Box of Text Technology", in: *Jaarboek Voor Nederlandse Boekgeschiedenis*, (Nijmegen: Vantilt 2013), pp. 201–204.

Adriaan van der Weel, *Changing Our Textual Minds : Towards a Digital Order of Knowledge*, (Manchester: Manchester University Press 2011).

Adriaan van der Weel, "New Mediums: New Perspectives on Knowledge Production", in: Wido van Peurzen, Ernst Thoutenhoofd, & Adriaan van der Weel (eds.), *Text Comparison and Digital Creativity*, (Leiden: Brill 2010).

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

Rene Wellek & Austen Warren, *Theory of Literature*, (Harmondsworth: Penguin Books 1963).

Marlo Welshons, *Our Cultural Commonwealth: The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences*, (2006), <http://msc.mellon.org/research-reports/Our Cultural Commonwealth.pdf/view>.

Helen Westgeest, "Visualizing Research and Visual Communication of Research", in: Helen Westgeest (ed.), *Making Research Visible to the World*, (Amstelveen: Canon Foundation in Europe 2010), pp. 9–14.

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).

Matthew Wilkens, "Canons, Close Reading, and the Evolution of Method", in: Matthew Gold (ed.), *Debates in the Digital Humanities*, (Minneapolis: University of Minnesota Press 2012).

Leland Wilkinson, *The Grammar of Graphics*, (New York: Springer 2005).

M.D. Wilson, "The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2", in: *Behavioural Research Methods, Instruments and Computers*, 20:1 (1988), pp. 6–11.

W.K. Wimsatt, *The Verbal Icon: Studies in the Meaning of Poetry*, (Lexington:

University Press of Kentucky 1954).

Langdon Winner, "Do Artefacts Have Politics?", in: *Daedalus*, 190:1. Modern Technology: Problem or Opportunity? (1980), pp. 121–136.

Ian H. Witten, "Text Mining", in: Munindar P . Singh (ed.), *The Practical Handbook of Internet Computing*, (Boca Raton; London: Chapman and Hall/CRC 1999), p. 198.

Susan Wittig, "The Computer and the Concept of Text", in: *Computers and the Humanities*, 11 (1978), pp. 211–215.

William Wulf, "The Collaboratory Opportunity", in: *Science*, 261:5123 (1993), pp. 854–855.

Sally Wyatt, "Technological Determinism Is Dead: Long Live Technological Determinism", in: *The Handbook of Science & Technology Studies*, (Cambridge: MIT Press 2008), pp. 165–180.

Case study

Primary sources

Seamus Heaney, "Feeling into Words", in: *Finders Keepers*, (London: Faber and Faber 2003).

Louis MacNeice, "Experiences with Images", in: Alan Heuser (ed.), *Selected Literary Criticism of Louis MacNeice,* (Oxford: Clarendon Press 1987)

Louis MacNeice, *Modern Poetry: A Personal Essay*, (Oxford: Clarendon Press 1968).

Louis MacNeice, *The Strings Are False: An Unfinished Autobiography*, (Oxford: Oxford University Press 1966).

Louis MacNeice, Peter McDonald (ed.), *Collected Poems*, (London: Faber & Faber 2007).

William Butler Yeats, Norman Jeffares (ed.), *Yeats's Poems*, (Basingstoke: Macmillan 1996).

Secondary sources

Terence Brown, *Louis MacNeice: Sceptical Vision*, (Dublin; New York: Gill and Macmillan 1975).

Richard Danson Brown, "Neutrality and Commitment: MacNeice, Yeats, Ireland and the Second World War", in: *Journal of Modern Literature*, 28:3 (2005), pp. 109–129.

Heather Clark, "Revising MacNeice", in: The Cambridge Quarterly, 31:1 (2002), pp. 77–92.

Neil Corcoran, "The Same Again? Repetition and Refrain in Louis MacNeice", in: *The Cambridge Quarterly*, 38:3 (4 August 2009), pp. 214–224, <http://dx.doi.org/10.1093/camqtly/bfp012>.

Alan Gillis, "“Any Dark Saying”: Louis MacNeice in the Nineteen Fifties", in: *Irish University Review*, 42:1 (20 May 2012), pp. 105–123, <http://dx.doi.org/10.3366/iur.2012.0011>.

John Goodby, "Louis MacNeice", in: David Scott Kastan (ed.), *The Oxford Encyclopedia of British Literature*, (Oxford: Oxford University Press).

Edna Longley, "Louis MacNeice: Aspects of His Aesthetic Theory and Practice", in: Jacqueline Genet & Wynne Hellegouarc'h (eds.), *Studies on Louis MacNeice*, (Caen: Presses universitaires de Caen 1988).

Edna Longley, *Louis MacNeice: A Study*, (London: Faber and Faber 1988).

Derek Mahon, "MacNeice in England and Ireland", in: Terence Brown & Alec Reid (eds.), *Time Was Away: The World of Louis MacNeice*, (Dublin: Dolmen Press 1974).

Peter McDonald, "The Falling Castle", in: Jacqueline Genet & Wynne Hellegouarc'h (eds.), *Studies on Louis MacNeice*, (Caen: Presses universitaires de Caen 1988).

Peter McDonald, "Louis MacNeice: The Burning Perch", in: Neil Roberts (ed.), *Companion to Twentieth-Century Poetry*, (Hoboken: Wiley-Blackwell 2003).

William T. McKinnon, "MacNeice's Pale Panther: An Exercise in Dream Logic", in: *Essays in Criticism*, XXIII:4 (1973), pp. 388–398, <http://eic.oxfordjournals.org/cgi/doi/10.1093/eic/XXIII.4.388>.

Simon Workman, "“To Be Tired of This Is to Be Tired of Life”: Louis MacNeice's London", in: *Irish Writing London: Volume 2*, Tom Herron, (London: Bloomsbury Academic).

Websites

CELT, <http://www.ucc.ie/celt/>

Computer Aided Textual Markup & Analysis, <http://www.catma.de>

Digital Research Tools Directory, <http://dirtdirectory.org>

Falcons, <http://ws.nju.edu.cn/falcons/objectsearch/index.jsp>

Friend of a Friend, <http://www.foaf-project.org>

Harvard General Inquirer, <http://www.wjh.harvard.edu/~inquirer>

Linguistic Inquiry and Worc Count, <http://www.liwc.net/> (12 June 2013)

Natural Language Toolkit, <http://www.nltk.org/>

Open Annotation Collaboration, <http://www.openannotation.org >

Stefanie Posavec, "Literary Organism", <http://www.stefanieposavec.co.uk>

"Research Data Management Guidance", <http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library/research-data-mgmt>.

The Rossetti Archive <http://www.rossettiarchive.org/>

The Algernon Charles Swinburne Project, <http://swinburnearchive.indiana.edu/>

TAPoR, <http://www.tapor.ca/>

"TEI Guidelines for Electronic Text Encoding and Interchange", <http://www.tei-c.org/Guidelines/P5/>

UCREL Semantic Analysis System, < http://ucrel.lancs.ac.uk/usas/>

W3C, "Extensible Markup Language (XML) 1.0 (Fifth Edition)", <http://www.w3.org/TR/2008/REC-xml-20081126/>.

Women Writers Project, <http://www.northeastern.edu/nulab/women-writers-project-2/>

World Wide Web Foundation, <http://webfoundation.org>

Appendices

Appendix A: Glossary of Technical and Statistical Terms

Euclidean distance The Euclidean distance between two points represents the general similarity of the vectors that are associated with these two points. To calculate the Euclidean distance, all the values for these two points need to be subtracted. These differences are then squared and summed. Euclidean distances can effectively be visualised using a dendrogram.

Chi-Squared test Chi-squared test is a statistical method which can be used to determine whether or not the variation within the values that are collected for a particular variable follows standardised probability rules. In the null hypothesis, there is no difference between the expected values and the observed values. If the result of the formula exceeds a given critical value, however, this null hypothesis must be rejected. A rejection implies that the variable which is investigated indeed has an effect on the values.

Cosine similarity Cosine similarity is a formula which can give an impression of the similarity between two vectors. This formula normally results in a value in between one and minus one, but if this formula is used to evaluate the similarity of two texts, based on their word frequencies, these frequencies can obviously not be negative. In this specific situation, the cosine similarity lies in between zero and one. If the two vectors are completely identical, the value is one. If the value is zero, all the values in these two vectors are dissimilar.

KWIC The acronym KWIC stand for “Keyword in Context List”. It is an operation in which a computer retrieves all the passages which contain a specific search term. These search results are then shown within their original context, consisting of a number of words before and after the occurrence of this search term. In many cases, users of KWIC software can specify the number of words or characters that are shown to clarify the context. In this thesis, the term is used synonymously with "concordance".

OAC

The *Open Annotation Collaboration* is a data model which represents the way in which scholars can annotate particular sources, or fragments within these sources. This model stipulates that annotations consist of a target, which is the source that is annotated, and a body, which refers to the comment which is made about this source. This model can be implemented using semantic web technologies. The main advantage of using such technologies is that the annotations are not tied exclusively to a single software environment, and that they can be shared more easily across different platforms and across different applications.

PCA

Principal Component Analysis is a form of multi-variate analysis in which a large number of variables can be replaced by a much smaller number of variables. The method is based on the calculation of the eigen vectors of the covariance matrix of all the data values. The method aims to create new variables which can account for most of the variability in the full data set. These new variables are referred to as the principal components. If the first two principal components account for most of the variability, the nature of the full data set can be clarified by plotting these two principal components.

Perl

The *Practical Extraction and Report Language* is a programming language, originally developed by the linguist Larry Wall. It offers extensive possibilities for regular expressions and for object-oriented programming.

Processing

Processing is a programming language, based on the Java language, which consists mainly of methods and classes for the creation of data visualisations. Processing is also a software environment in which Processing code can be executed.

Naive Bayes Naive Bayes is a supervised machine learning algorithm. It often starts with a process in which human beings train software applications by manually supplying categories for the data in a training set. The application subsequently develops a model on the basis of these labelled data. This model can be used to make predictions about new unlabeled data. Naïve Bayes is based on Bayes' theorem, which describes the probability of an event, based on occurrences of events that are related.

N-gram An n-gram can be described generally as a sequence of textual units. These "grams", or textual units, are usually words or characters. More rarely, they can also be syllables or phonemes. The "n" in this term refers to the length of the sequence. A bigram (or a two-gram), for instance, can consist of a sequence of two words or of two characters. Analyses of bigrams or trigrams can be useful in applications which focus on occurrences of specific phrases.

R R is both a software environment and a programming language. The software environment can be used to perform statistical analyses and to produce graphics for the clarification of such analyses. R offers support for a wide range of standard statistical operations, and its functionalities can be extended considerably through the installation of additional packages. R was developed at Bell Laboratories, and it is available as free software under a GNU Public License.

RDF The *Resource Description Framework* provides a generic model for the description of resources. It envisages descriptions or annotations broadly as assertions consisting of three parts: a subject, a predicate and an object. The assumption is that any statement can be expressed using this tripartite structure. RDF has emerged as a standard model for the exchange of data on the web. RDF triples are often visualised as graphs.

Standard deviation The standard deviation is a statistical measure, which can give an indication of the degree of variation within the values of a dataset. Standard deviations are calculated by taking the square root of the variance of all the values. The variance, in turn, is produced by calculating the average of the squared differences between all values and their mean. If the standard deviation is low, this means that all values are close to this mean value. A high value means that the values are dispersed more widely. A standard deviation of one means that the values are distributed according to a standard normal distribution, and that a plot of these values results in a bell curve.

Td-idf

The abbreviation td-idf stands for frequency-inverse document frequency. It is a statistical operation which was designed to indicate the importance of a specific term within the context of a corpus. The td-idf formula assigns a high weight to a term if it occurs in a small number of documents. Terms which occur in all documents will receive the value zero. This is generally the case for very frequent terms, such as articles, prepositions or pronouns. The formula can thus be use to retrieve the rarer, more distinctive words from a corpus.

Token

Tokens result from the process of tokenisation. In this latter process, a full text is divided into its constituent units. The aim of tokenisation is often to separate a text into individual words. In this situation, the term “tokens” refers to the total number of words in a text.

Topic Modelling

The type of searches that are enabled by search engines or by library catalogues are generally based on the assumption that people know beforehand what they are searching for, and that they can supply specific search terms. Topic Modelling is a fundamentally different approach, in which algorithms organise the textual data by themselves, and in which they try to extract some of the topics that are discussed in a corpus. Topic Modelling is based on the Latent Dirichlet Allocation model, which was first developed by David Blei. The model considers the frequencies of all the words that occur in the corpus, and combines these with data about the documents in which these words are used. On the basis of these data, Topic Modelling algorithms can produce lists of words which are often used in combination, and which can be assumed to refer to the same topic. These clusters of words are initially unlabelled, and researchers who use Topic Modelling must interpret and label these groups of words themselves. These word clusters can give a rough indication of the topics that are discussed within a collection of text documents.

Type Types are the distinct words that occur in a text. Types are often associated with frequencies, which reflect the number of times the type is used. Data about the number of types and the number of tokens can be used to calculate the type-token ratio, which gives an impression of the diversity of the vocabulary.

XML The *eXtensible Markup Language* can be used to provide explicit descriptions of specific aspects of text fragments through the addition of inline encoding. As is also manifest in the final two letters of the name of the standard, XML is used to ‘mark up’ specific aspects of a textual document. Marking up a text entails two things: (1) selecting or situating a certain logical or structural component within the text and (2) giving information about the fragment which is selected. The descriptive terms which are added to the document are referred to as *elements*. The elements which are allowed in a particular XML-based encoding language are listed in a *Document Type Definition* or in a *Schema*.

z-score The z-score of a value in a data set expresses its distance to the mean of this data set. This distance is expressed in standard deviations. They provide an intuitive indication of the position of an individual value within the context of the full collections of values. A negative value means that the value is below the mean, and a positive value indicates that this value is higher than the mean. When all the values in a dataset are converted to z-scores, the mean of these values will be zero, and the standard deviation of this collection of z-scores will have value one. An important advantage of z-scores is that they do not have a unit of measurement. Data sets with different distributions and with different units of measurement can be compared effectively by firstly converting all their values to z-scores.

Appendix B: Ontology of Literary Terms

Devices based on repetitions of sounds	Alliteration		
	Assonance		
	Consonance		
	Internal Consonance Rhyme		
Devices based on changes in meaning	Metaphor		
	Simile		
	Metonomy		
	Personification		
	Synechdoche		
Devices based on repetition or word Order	Paronymy		
	Chiasmus		
	Anaphora		
Prosodic Techniques	Rhyme	Perfect Rhyme	
		Slant Rhyme	Assonance Rhyme
			Consonance Rhyme
		Semi Rhyme	
		Deibhide rhyme	
		Aicill Rhyme	
		Internal Rhyme	
	Rhythm		
	Metre	Iambic	
		Trochaic	
		Spondaic	
		Dactylic	
		Trimeter	
		Tetrameter	
		Pentameter	
		Hexameter	
Form	Couplet		
	Three line form	Tercet	
		Triplet	
	Quatrain		
	Quintain		

	Sestet	
	Fourteen line form	Sonnet
Texture		
Diction		
Syntax		
Volume		
Tone		
Mood		
Structural terms	Poem	
	Stanza	
	Line	

This ontology is not intended as an exhaustive list of all existing literary techniques. It concentrates principally on the terms which were discussed in Chapter 2 of this dissertation, and on those terms which were relevant for the case study that was conducted for this dissertation.

Samenvatting

De mogelijkheden en de beperkingen van algoritmische literatuurkritiek

Computers zijn in essentie machines die kunnen rekenen, maar omdat specifieke numerieke codes geassocieerd kunnen worden met letters en met woorden zijn computers ook in staat om teksten in natuurlijke talen te bewerken en te analyseren. De technologieën die we kunnen gebruiken voor het doorzoeken van teksten worden steeds geavanceerder, en dit heeft als gevolg dat we meer en meer aspecten van het menselijke leesproces aan de computer kunnen uitbesteden. De talloze grootschalige digitaliseringsprojecten die in de afgelopen decennia zijn uitgevoerd hebben bovendien geresulteerd in een enorme hoeveelheid aan machine-leesbare teksten. Verschillende onderzoekers hebben aangetoond dat deze ontwikkelingen belangrijke gevolgen kunnen hebben voor de literatuurwetenschap. In zijn boek *Macroanalysis* beschrijft Matthew Jockers, bijvoorbeeld, dat computeralgoritmes kunnen worden toegepast op corpora van honderden of zelf duizenden teksten, en dat het hierdoor mogelijk wordt om onderzoek te doen naar brede historische ontwikkelingen rond het gebruik van bepaalde stijlfiguren, of rond de opkomst en de ondergang van literaire genres. Deze methodiek, waarin computers op zoek gaan naar patronen in grote tekst corpora, wordt momenteel vaak aangeduid met de door Franco Moretti geïntroduceerde term *distant reading*. Binnen de literatuurwetenschap blijft het gebruik van digitale technologieën momenteel echter nog beperkt tot een kleine groep pioniers. Het merendeel van het literatuuronderzoek vindt nog plaats via het traditionele hermeneutische model, waarin de meest relevante teksten grondig worden bestudeerd via de *close reading* methode. Een belangrijk doel van mijn proefschrift was om de mogelijkheden en de beperkingen van computergebaseerd literatuuronderzoek duidelijk in kaart te brengen, en, met name, om de belangrijkste verschillen en overeenkomsten vast te stellen tussen *distant reading* en *close reading*.

Om deze vraag te kunnen beantwoorden heb ik onder meer een gedetailleerde beschrijving gemaakt van de onderzoeksmethoden die binnen de computationele literatuurwetenschap worden gebruikt. Over het algemeen beginnen computationele analyses met een proces waarbij de oorspronkelijke teksten worden geconverteerd naar discrete data. Er kunnen verschillende typen van data worden onderscheiden. Veel studies baseren hun analyses op de frequenties van de woorden die voorkomen in een tekst. Wanneer de focus van een onderzoeker hoofdzakelijk ligt op concepten, en minder op specifieke woordvormen, kan het nuttig zijn om gebruik te maken van lemmatiseringssoftware, die alle woorden die vervoegd of verbogen zijn kunnen terugbrengen naar een basisvorm. Door gebruik te maken van *Part of Speech* taggers kunnen er gegevens worden verzameld over de grammaticale en syntactische categorieën van de gevonden woorden. Via technieken zoals *Semantic Tagging* of *Topic Modelling* kunnen onderzoekers zich een beeld vormen van de onderwerpen die in tekst worden besproken. Verschillende studies hebben echter aangetoond dat technologieën voor semantische analyses vaak nog geen goede resultaten opleveren voor literaire teksten. Deze verschillende data stellen onderzoekers in staat om de stijl van teksten op een zeer systematische manier te beschrijven. Deze stilistische kenmerken kunnen vervolgens worden ingezet bij het vergelijken van teksten uit verschillende genres, van verschillende auteurs, of uit verschillende historische periodes. Gegevens over woordfrequenties zijn vaak ook zeer effectief bij het dateren van teksten, of bij het vaststellen van de waarschijnlijke auteurs van teksten.

Het onderzoek dat gebruik maakt van digitale onderzoeksinstrumenten verschilt op een aantal punten van de meer conventionele literatuurkritiek. Een eerste verschil is dat computationele analyses vaak gebaseerd zijn op gegevens over de taalkundige aspecten van de teksten. Digitale onderzoeksinstrumenten bieden nog vrijwel geen ondersteuning voor de herkenning van de meer specifieke aspecten die veel aandacht krijgen in traditioneel literatuuronderzoek, zoals stijlfiguren, de connotaties van specifieke woorden, vormen van beeldspraak en literaire thema's. Een tweede verschil is dat computers teksten op een non-responsieve en een context-onafhankelijke manier analyseren. Wanneer mensen literatuur lezen heeft hun kennis van de sociale of historische context meestal een invloed op hoe woorden worden geïnterpreteerd en beoordeeld. Traditionele literatuuranalyses richten zich bovendien vaak op de manier waarop specifieke woorden en stijlfiguren worden gecombineerd op het niveau van zinnen of alinea's. Bij het samenstellen van kwantitatieve data over teksten gaat de oorspronkelijke context van de fenomenen die worden geanalyseerd meestal verloren. Mede hierdoor is het lastig om responsieve tekstanalyse-software te ontwikkelen. De regels die zijn vastgelegd in een algoritme worden momenteel vaak met een onbuigzame consistentie toegepast op alle teksten in het corpus. Op de derde plaats is van belang om vast te stellen dat kwantitatieve analyses van data over taalkundige aspecten zelf geen interpretatie toevoegen. De regelmatigheden, verbanden of uitzonderingen die via computers kunnen worden vastgesteld kunnen de premissen leveren voor een conclusie, maar er is nog steeds een menselijke onderzoeker nodig om deze conclusies te formuleren. Vanwege de focus op waarneembare taalkundige aspecten van teksten worden digitale onderzoeksinstrumenten vaak ingezet voor onderzoeksvragen op het gebied van de literatuurgeschiedenis, die vaak met een enkelvoudig antwoord kunnen worden opgelost. De bestaande digitale methoden worden minder vaak gebruikt voor het verruimen of het veranderen van de manier waarop literaire teksten kunnen worden geïnterpreteerd.

Als onderdeel van mijn proefschift heb ik een experimentele studie uitgevoerd waarin ik heb onderzocht of een aantal van de genoemde beperkingen deels of volledig kunnen worden weggenomen. Meer specifiek heb ik onderzoek gedaan naar de vraag of de stijlfiguren die in traditioneel onderzoek worden bestudeerd ook door computers kunnen worden herkend. Indien het inderdaad mogelijk is om dit soort data te genereren, is het op de tweede plaats ook van belang om te weten of deze nieuwe data mogelijk ook kunnen van belang kunnen zijn bij de interpretatie van teksten. Beide vragen heb ik onderzocht via een *case study*, die zich richtte zich op de poëzie van de Noord-Ierse dichter Louis MacNeice. In het kader van de case study heb ik een groot aantal algoritmes ontwikkeld voor de herkenning van een aantal stijlfiguren. Omdat veel literaire technieken gebaseerd zijn op specifieke combinaties van klanken, heb ik bij het begin van de studie een methode ontwikkeld voor het maken van fonetische transcripties van alle versregels. Aan de hand van deze fonetische transcripties heb ik vervolgens een aantal regels gedefinieerd waarmee verschillende vormen van rijm kunnen worden herkend. Eindrijm, bijvoorbeeld, houdt in dat er een exacte herhaling is van de fonemen aan het eind van een versregel. Op dezelfde manier heb ik algoritmes ontwikkeld waarmee stijlfiguren zoals alliteratie, binnenrijm en consonantie kunnen worden gekwantificeerd. Een belangrijke uitdaging bij de ontwikkeling van software voor de herkenning van stijlfiguren is dat er veel variatie bestaat in de manier waarop deze concreet zijn toegepast. De gedichten van MacNeice bevatten veel herhalingen en combinaties van klanken, maar het is niet altijd eenvoudig om dit soort herhalingen ook eenduidig te categoriseren. Wanneer de regels voor het herkennen van, bijvoorbeeld, eindrijm worden versoepeld levert dit naast een aantal aanvullende resultaten vaak ook veel valse positieven op. Een ander probleem is dat de precieze eigenschappen van een stijlfiguur niet altijd in een algoritme kunnen worden vastgelegd. Dit probleem speelde met name bij de algoritmes voor het vinden van literaire allusies. Zoals ook is aangetoond in een aantal eerdere studies kunnen de exacte verschillen tussen een echte literaire allusie en zinnen die simpelweg een aantal woorden delen lastig in vaste regels worden gevangen.

De data die in de case study zijn verzameld kunnen worden gebruikt om een aantal grotere patronen binnen het oeuvre van MacNeice bloot te leggen. Over de bundels die over het algemeen het minste worden gewaardeerd door critici, *Autumn Sequel* en *Ten Burnt Offerings*, kon worden vastgesteld dat de frequenties van de onderzochte stijlfiguren ook lagere standaarddeviaties hebben. De gedichten in deze bundels hebben een lagere *type-token ratio*, en het gemiddeld aantal lettergrepen per woord ligt ook lager. Uit een analyse van de beeldspraak bleek, zoals ook is vastgesteld door eerdere critici van het werk van MacNeice, dat de poëzie veel verwijzingen bevat naar religie, naar de zee en naar vervoersmiddelen zoals treinen en auto's. Uit de computationele analyse bleek eveneens dat er ook veel woorden zijn gebruikt die associaties oproepen met geld, met oorlog en met voedsel. Deze laatste vormen van beeldspraak zijn nog niet eerder besproken in de bestaande literatuur over MacNeice. Vaak kan het ook interessant zijn om te onderzoeken of de gedichten die specifieke stilistische kenmerken delen ook thematische overeenkomsten vertonen. Gedichten waarin veel alliteratie voorkomt beschrijven vaak een drang naar escapisme. Het Keltische deibhide rijm wordt vaak gebruikt in gedichten waarin sociale verschillen tussen personen of tussen groepen van personen worden benadrukt, en het gebruik van semi-rijm hangt vaak samen met verwijzingen naar de chaos en de onpersoonlijkheid van de moderne consumptiemaatschappij. Op basis van de fonetische transcripties van alle versregels kon bovendien worden vastgesteld dat de gedichten die bijna uitsluitend vrouwelijk eindrijm bevatten eveneens eilanden als metafoor gebruiken.

In het laatste hoofdstuk van het proefschrift worden de belangrijkste verschillen en overeenkomsten besproken tussen de twee benaderingen in de literatuurkritiek. Een belangrijk verschil is dat het gebruik van digitale onderzoeksinstrumenten een zeer praktische vorm van onderzoek inhoudt, waarbij er nieuwe ideeën worden ontwikkeld via de bouw van non-tekstuele of non-lineaire resultaten. Via computers kunnen onderzoekers bovendien ook andersoortige observaties doen. Computationele methoden stellen onderzoekers in staat om aspecten van teksten bloot te leggen die niet of minder gemakkelijk kunnen worden waargenomen door menselijke lezers. Zaken zoals de *type-token ratio*, of het gemiddeld aantal lettergrepen per woord kunnen, in het geval van grote corpora, alleen door computers worden berekend. Hetzelfde geldt vaak voor patronen in het gebruik van woordsoorten die heel frequent zijn, zoals lidwoorden of voorzetsels. Voor menselijke lezers is het vaak ook lastig om vast te kunnen stellen dat bepaalde fenomenen afwezig zijn. In de case study is bijvoorbeeld opgemerkt dat er in een aantal gedichten van MacNeice geen consonantie voorkomt, terwijl dit stijlfiguur juist wel frequent is in het merendeel van de teksten. Zonder computers zijn dit soort observaties lastig te maken. Wanneer de patronen die door computers worden vastgesteld stroken met bestaande inzichten kan dat waardevol zijn, omdat deze aanvullende observaties de bestaande argumenten dan op een andere manier kunnen onderbouwen. Het omgekeerde geval, waarin de gegenereerde patronen in tegenspraak zijn met bestaande ideeën kan eveneens nuttig zijn, omdat de pogingen om deze verschillen te verklaren vaak kunnen leiden tot een beter begrip van de bestudeerde teksten.

Naast deze verschillen zijn er ook een aantal overeenkomsten. Hoewel er soms wordt beweerd dat het gebruik van digitale methodes ook een overgang impliceert naar een meer objectieve benadering, is het belangrijk om vast te stellen dat de functionaliteiten van onderzoeksinstrumenten vaak op subjectieve beslissingen zijn gebaseerd. Onderzoeksinstrumenten worden ontwikkeld door menselijke programmeurs die bewust of onbewust beslissingen moeten nemen over het soort resultaten dat de tools moeten opleveren. Tools worden ontwikkeld binnen een bepaalde wetenschappelijke en methodologische traditie en voor een heel duidelijk afgebakend doel. De data die door dit soort tools worden gegenereerd zijn niet noodzakelijk objectiever dan de gegevens die handmatig door menselijke wetenschappers worden verzameld.

Een tweede overeenkomst is dat er binnen beide benaderingen zowel inductieve als deductieve methoden worden gebruikt. De literatuurwetenschapper Stanley Fish betoogde dat het gebruik van algoritmes noodzakelijkerwijs een inductieve onderzoeksmethode inhoudt, waarin losse observaties pas na een experiment kunnen worden gekoppeld aan een theorie die deze observaties kan verklaren. Deze kritiek is echter niet terecht. Op de eerste plaats zijn inductieve methoden ook relevant voor het traditionele onderzoek. Wanneer een criticus een nieuw boek begint te lezen is er vaak nog geen volledig uitgewerkte hypothese. Hiernaast is het ook zo dat computergestuurd onderzoek deductief van aard kan zijn. Hoewel de uitkomst van experimenten soms zeer onvoorspelbaar kunnen zijn, kunnen de experimenten wel worden uitgevoerd met het doel om bestaande theorieën te staven of te weerleggen.

Computers zijn van belang voor de literatuurwetenschap omdat ze onderzoekers in staat stellen om teksten te bestuderen vanuit nieuwe, vaak verassende, perspectieven. Hierbij is het van belang om te onderstrepen dat computationele methoden geen doel op zich vormen. Literatuurwetenschappers die experimenteren met de mogelijkheden van digitale technologieën moeten kritisch blijven nadenken over de vraag of deze methoden ook daadwerkelijk inzichten opleveren die relevant zijn voor de literatuurwetenschap.

Acknowledgements

Writing this dissertation has been a very enjoyable and a deeply rewarding experience. On the whole, and despite some occasional difficulties, it felt like a great privilege to be able to develop my knowledge about a topic which fascinates me profoundly. Pursuing a PhD degree unavoidably implies many hours of solitary work, which, in my case, were mostly spent reading, writing or programming. To a great extent, the motivation and the enthusiasm that I felt during all (or most) of these activities were born out of the gratifying and humbling feeling of being supported by many different people who advised me, encouraged me or otherwise inspired me.

First and foremost, I owe an immense debt of gratitude to my promotor Adriaan van der Weel. When I was just a regular humanities student, roughly around the turn of the millennium, he first introduced me to the miraculous world of digital texts. His lectures sparked a fascination which has stayed with me ever since. He has been a very inspiring, a very patient and a very conscientious supervisor, who motivated me in many different ways to get the most out of myself. Our many discussions about the topics of this thesis have strongly sharpened my critical thinking, and have contributed enormously to my overall academic development.

I want to extend my gratitude to Paul Hoftijzer, Fleur Praal and Erik Kwakkel, my other colleagues at the MA Book and Digital Media Studies, for their support and for their continuous interest in my dissertation. When it became clear, at the beginning of 2014, that our research institution LUCAS had made available a number of grants for external PhD candidates, Paul made sure that some of this funding could be used to free me from my teaching duties in the fall of 2014. During this period, I could work on my dissertation in a much more focused manner, and my research made an enormous progress because of this. In a similar fashion, Fleur was always willing to take over some lecturing duties or correction work from me whenever the work on my dissertation was hectic. Preparing classes and teaching classes together with Fleur has always been a great pleasure, and the many debates that we have had, about digital scholarship or about the changing ways of reading, were often useful directly for the argumentation that I was constructing in my thesis.

I am thankful also to my second promotor, Karina van Dalen-Oskam. After we had discovered that the revised PhD regulations of Leiden University contained rules which stipulated that all PhD students must be assigned two *promotores*, she gladly accepted the invitation to take on this responsibility. The fact that she had not been involved in some of the earliest stages of my promotion can perhaps be considered a disadvantage, but this was clearly balanced out by the simultaneous benefit that she could evaluate the thesis from an entirely fresh perspective. The meticulous and insightful feedback that I received from her greatly helped me to improve the contents and the structure of this dissertation.

I am grateful to Korrie Korevaart, Viola Stoops, Thony Visser and Geert Warnar at the LUCAS institute for facilitating my research and for assisting me with all kinds of practical matters on numerous occasions.

I have formally written this thesis as an external PhD candidate, which meant that I mainly had to carry out the research in my spare time, next to my work as a university teacher and as a project manager at Leiden University Libraries. The combination of these activities was occasionally rather overwhelming, and at times I certainly wished, to quote MacNeice, that I could make time elastic. These practical challenges were forcibly attenuated, however, by the fact there were many interesting and relevant interconnections between my daytime job and the work on my dissertation. While this is perhaps most obvious for my work for Book and Digital Media Studies, there were also many fruitful opportunities for cross-fertilisation between my research and my work for the university library. The projects that I have been involved in enabled me to develop my knowledge about a wide range of topics which are connected to my research, including digital scholarship, open access publishing, text mining, data mining and research data management. I feel grateful to all the close colleagues I have had to pleasure to work with, and in particular to Fieke Schoots, Isabel Brouwer, Saskia van Bergen, Cynthia van der Brugge, Lucas van Schaik, Birte Kristiansen, Kurt De Belder, Kees Konings and Bas Vat. Specifically, I want to thank Laurents Sesink, who has been my manager during the last two years. He stimulated me to reflect critically on some of the practical challenges surrounding digital scholarship, and he also alerted me to many new and exciting developments. Importantly, he understood that the final stages of my research project were very time-consuming, and he generously allowed me to work on my thesis occasionally, at moments when I should actually be doing work for the library.

I am sincerely thankful to my fantastic brothers Erik and Marcel, not only for their willingness to act as *paranimfen* during my PhD defense, but also because of their unwavering interest in my thesis, and because of their valuable comments on preliminary drafts of this thesis. I also wish to express my deep gratitude to my parents, Riet and Frans. They have always encouraged me to pursue a PhD degree and have been warmly supportive ever since I started this research project. I want to thank them both for the countless times they took care of my children, and I specifically want to thank my mother for letting me stay in her woodside home during some of her holidays. Such days, on which I could concentrate on my thesis in isolation and without any distractions, were all highly productive, next to being very pleasant.

I dedicate this dissertation to Mariëlle, Eline and Lars. Mariëlle has supported me, listened to me and believed in me throughout the entire writing process, and I am deeply indebted to her for making it possible for me to work on my dissertation, next to our hectic family life. She is a great mother to our two beautiful children, who were both born while I was working on this dissertation. Looking back, both at becoming a father and at writing a dissertation, it is clear to me that there are strong similarities between these two activities. As is the case for fatherhood, I now realise that I did not have a realistic image of what it would be like to pursue a PhD degree before I actually began this process. Both endeavours have dominated my personal life in the last five years, and they often made the very notion of having some spare time to myself seem like a distant memory from the past. Like writing a PhD, raising children demands patience, dedication, forbearance and perseverance, but ultimately, the sense of accomplishment, the elementary happiness and the immense feeling of pride that can result from watching the stages of development in both of these areas is indescribable and immeasurable.

Curriculum Vitae

Peter Verhaar (1976) studied *English Language and Literature* at Leiden University. During his studies, he spent one year abroad, at University College Dublin in Ireland. He graduated cum laude in 2002, with a master's thesis about the plays of the Northern-Irish dramatist Brian Friel. After his graduation, he briefly worked at the Peace Palace Library in The Hague, and, from 2003 to 2005, he was employed by IDC Publishers in Leiden as a metadata specialist. During this same period, he worked as teacher of Academic Writing in English at the University of Amsterdam, and he also studied computer science as a part-time student at the Open University. In 2005, he began to work as a project manager at Leiden University Libraries. Initially, his focus was mostly on providing digital access to the library's extensive special collections. After 2008, his focus shifted to projects in the fields of research support and digital scholarship. Since 2004, he has also worked as a university lecturer at the MA programme Book and Digital Media Studies at Leiden University. He has taught several courses about text encoding, database theory, the digital humanities and media theory, among many other topics. In the fall of 2010, he began to work on a dissertation as an external PhD candidate, at the Leiden University Centre for the Arts in Society.