

Novel Disruption:
How Text Mining May Change Literature

Master's Thesis
Book and Digital Media Studies

Loren Snel

Table of Contents

Word of Thanks

Introduction

Chapter 1: Text Mining and its Introduction to Trade Fiction Publishing

- 1.1. Technological determinism
- 1.2. What is text mining?
 - 1.2.1. A superior form of reading
 - 1.2.2. An academic field
- 1.3. How does text mining work?
- 1.4. Deconstructing misunderstandings about text mining

Chapter 2: Text Mining as a Publishing Tool: Three Case Studies

- 2.1. *No longer Lost in Translation: Text mining and Acquisition of Foreign Titles*
 - 2.1.1. Acquisition and commission
 - 2.1.2. Foreign manuscript scanner
 - 2.1.3. Risks and limitations
- 2.2. *The Tumbling of Editing's Ivory Tower: Text mining and Marketing*
 - 2.2.1. Marketing
 - 2.2.2. Text and data mining for clues
 - 2.2.3. Risks and limitations
- 2.3. *Strengthening Metadata: Text Mining and Sales*
 - 2.3.1. Sales and metadata
 - 2.3.2. Text mining for metadata
 - 2.3.3. Risk and Limitations

Conclusion

Bibliography

Word of Thanks

Pity, that it is not an academic custom to mention those in your thesis bibliography who helped you most in writing the thing. I have many to thank, and much to be thankful for.

First of all, my supervisor professor Adriaan van der Weel, who righteously warned me to avoid sentence structures like the one I just used. I still remember reading one of professor Van der Weel's engaging papers during my final bachelor semester, and realising I had found my master program. I thank him for his patience, for his support in the development of my ideas, and for inspiring me with his own.

I would also like to thank my loving family for their help, love and care, especially in the last two weeks of writing. This thesis would not have been half as adequate without their espresso's, warming meals and hugs. Dad, sis, but especially my mum, Lucy van Rooij, gets my deepest thanks for her editing assistance, and her otherworldly levels of patience every time I neared break-down. I can only dream of being as brilliant as she.

Furthermore, I would like to thank the publicists and book professionals who took the time to talk to me about tech in books. My thanks go to Gwena, Simon, Lisanne, Caroline and the people at Bookarang. It is good to know that those advancing such a vulnerable industry still make time to think about their work. I hope to be joining the ranks soon.

May this thesis be as entertaining, but not as straining, to read as it was to write.

Many thanks, or how does one end these things,

Loren Snel

Introduction

For publishing professionals and academics alike, keeping up with how technology affects literature seems a Sisyphus ordeal.

Technology and its potential applications are, if one is to believe TED talkers, technicians and trend watchers, evolving exponentially.¹ The continuous failings of both book academia and business in anticipating which supposedly ground-breaking possibilities will come true render the mission rather disheartening.² That is why, according to Simone Murray, the question of how technology impacts literature gets less and less attention with every unfulfilled prophesy. Hypertexts have not become the linearity-defying next big literary form. E-books have not removed paper from the centuries-old picture.³ And, in addition to Murray's list, the efficiency and power of Amazon and co have not wiped out the physical bookstore, yet.

Trumped prophesies on tech's influence may be taken up by some publishers as justification for ignoring new developments.⁴ Scholarship, however, does not throw in the sponge when predictions turn out to be harder to get right than expected. Moreover, an encouragement Murray ignores, new

¹ W. de Rek, 'Cijfers En Letteren', *Sir Edmund*, no. 38, 23 December 2017, p. 13.

² S. Murray, 'Charting the Digital Literary Sphere', *Contemporary Literature*, 56 no. 2 (2015), p. 311.

³ A. Preston, 'How Real Books Have Trumped Ebooks', *The Guardian*, Guardian News and Media, 14 May 2017 (8 June 2018).

⁴ As references to interviews with multiple people from the book publishing industry will make clear, specialist publishers especially seem to nurse an 'I'll believe it when I see it' mentality when it comes to possible new impact of tech on literature.

technological developments such as hypertext, e-books and print-on-demand have made their marks, just not in the exact ways prophesized. What Book and Publishing Studies need, therefore, is a willingness to research, predict and be falsified again and again, in Popperian fashion, combined with the determination to keep it up. For a new, and complex technology is already announcing itself on the horizon.

So far, Publishing Studies scholars have contributed to the debate on technology's effects on literature foremost by focusing on changes in fiction transmission. On the one hand, they have studied new technological substrates of fiction, such as desktops and e-books, and how these change reading habits, as well as revenue streams in the publishing business.⁵ On the other hand, scholars have studied how technologies like the Internet have led to the appearance and disappearance of fiction-selling stakeholders, such as Amazon and bookshops respectively. Publishing Studies have thus mainly emphasised technology's impact on literature's peripheries. The emergent technology at hand, however, is able to affect not just the means of communication of literature, but also its actual form and content. This thesis will discuss how exactly that new technology has that power.

Its name is text mining. Text mining is a technology originally created by and for an interdisciplinary academic field that manages to exploit texts as a source of extractable data using

⁵ T. S. Schweizer, 'Managing interactions between technological and stylistic innovation in the media industries', *Technology Analysis & Strategic Management*, 15 no. 1 (2002).

algorithmic analysis.⁶ For some, this description may ring some bells. Was it not a form of text mining that helped demask J. K. Rowling as the writer behind the pseudonym of Robert Galbraith?⁷ It was. Did the scientist who proved Italian author Domenico Starnone's writing style to be a ringer for Ferrante's, the author of world famous novels, not use text mining?⁸ Indeed he did.

The famous cases of Galbraith and Ferrante are indicative of the evermore prevalent application of text mining to fiction. The technology has already been proposed as a new mode of enquiry within Literary Criticism.⁹ Furthermore, the Digital Humanities have been analysing how text mining is created and applied within the academic fields that invented text mining.¹⁰ The technology has, however, not yet been looked at by the field of Publishing Studies. Now that trade fiction publishers are starting to take it up, study into the commercial application of text mining has become relevant if not necessary.

The start of the phenomenon can be somewhat traced back to one book. In 2016, Digital Humanities scholar Matthew L. Jockers and publicist Jodie Archer published *The Bestseller Code*.¹¹ In this non-fiction trade book, Archer and Jockers describe

⁶ D. Meyer et al, 'Text Mining Infrastructure in R.' *Journal of Statistical Software*, 25 no. 5 (2008).

⁷ P. Juola, 'Industrial uses for authorship analysis', *Mathematics and computers in sciences and industry*, INASE (2015).

⁸ E. ETTY, 'Ferrante, who cares?', *De Groene Amsterdammer*, no. 39, 27 September 2017 (9 June 2018).

⁹ M. G. Kirschenbaum, 'The remaking of reading: Data mining and the digital humanities', *The National Science Foundation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation*, (Baltimore: MD, 2007).

¹⁰ The Digital Humanities Quarterly journal is an excellent resource for research on text mining in the Humanities.

¹¹ J. Archer and M. L. Jockers, *The Bestseller Code: Anatomy of the Blockbuster Novel*, (St. Martin's Press, 2016).

how they developed an algorithm-based computer engine that could predict, with an over eighty-percent success rate, which novels from a corpus of 20,000 made it on to the *New York Times Bestsellers* list and which did not. Rather than giving away the technical details on how their engine worked, Jockers and Archer mostly focused on what it proved, especially addressing both established and aspirant authors. However, Archer and Jockers mainly raised the interest of publishers. Both Dutch start-up Bookarang, a book recommendation software service, and the Research & Development (R&D) department of Dutch publishing concern WPG¹² mention Archer and Jockers as their inspiration.¹³ Both companies apply text mining to literature, wishing to make a profit with it. They will not stand alone for long.¹⁴

Predicting publishers' implementation of new technology is a challenging task. Publishing Studies is a young field and has only recently taken it on.¹⁵ Early research by Schweizer suggested that the likelihood of technology being integrated differs per type of trade fiction publisher. Some suggest generalist publishers – commercial trade publishers with access to R&D resources – are more likely to welcome new technology into their business. The beforementioned WPG concern is, of course, a prime example of that theory. Others say smaller, more specialist and culturally

¹² L. Snel, Interview with Gwena Jouen of WPG/Schwung, 13 April 2018.

¹³ Bookarang was already founded in 2013, three years before *The Bestseller Code* came out, meaning the founders already knew about the possibilities of text mining literature before Archer and Jockers told them about their findings.

¹⁴ H. Chin-A-Fo, T. Jaeger, and one correspondent, 'Big data en algoritmes moeten worstsellers gaan voorkomen', *NRC*, 23 March 2017 (1 July 2018).

¹⁵ T. S. Schweizer, 'Multimedia Giants, Literary Publishers and New Technologies: Can Culture and Business Benefit from the Change of Rules in the Book World?', *International Journal of Arts Management*, 3 no. 3 (spring 2001), p. 51.

focussed publishers are the most likely to go for technological innovation in order to maintain a competitive edge.¹⁶ Simon Dikker Hupkes, editor at Dutch trade publisher Atlas Contact, also imagines smaller and more commercial publishers might take an interest in text mining functionalities.¹⁷

Although still emergent, text mining technology clearly demands the opening up of a new strand of research within Publishing Studies. Keeping track of how such developments affect fiction publishing and thereby literature is hard, especially when that technology is still so fresh. Yet, if we want publishing to keep some level of insight into the impact of technology on literature, it is important. Publishing Studies research can help attain some of that insight, something this thesis aims to contribute to. Although it is not Publishing Studies' task to shield the shaping of fiction from new influences, it does carry the responsibility to point out those influences.

Another objective of this thesis is that studying the effect of text mining on fiction can help unite Publishing Studies with Literary Studies. Literary Studies have often trivialised the effect of technology on content. Their concern, so they say, is the art of literature and what, through its analysis, that art can tell us. Consequently, they often risk dismissing the fact that their objects of study are also mere published works of fiction, products of a both pliant and formative publishing process.¹⁸ Publishing Studies

¹⁶ T. S. Schweizer, 'Multimedia Giants, Literary Publishers and New Technologies: Can Culture and Business Benefit from the Change of Rules in the Book World?', p. 56.

¹⁷ L. Snel, Interview with Simon Dikker Hupkes at Atlas Contact, 15 March 2018.

¹⁸ Literary Studies' long tradition of Barthesian scholarship could be a cause of this neglect. Once the author, the primary creator of a work, is declared

often recognises the intertwinement of the technology-enabled publishing process and the content it produces. In the recently published *The Written World*, professor of Comparative Literature and English at Harvard, Martin Puchner underpins Publishing Studies' philosophy by showing how, through time, different writing implements and substrates moulded the very writing they created and contained.¹⁹ Hopefully, the new phenomenon of text mining being used in the creation of fiction, can serve as an opportunity for Publishing Studies and Literary Studies to unite. Puchner's new work already serves as an exemplary step in the right direction.

It may be clear by now that this thesis claims that text mining has the power to change literature. Yet how exactly does it have that power? As we shall prove and argue, text mining is able to change literature by means of its integration into specific business operations of trade fiction publishers.

It is important to note that, although the work of WPG and Bookarang serve as indicators, text mining is at this point far from being commercially implemented within the book market. WPG is momentarily developing text mining-based products intended for different stakeholders in the book market, including their own trade fiction publishing imprints.²⁰ Bookarang is not a publishing concern, but uses novel texts sent to them by publishers to offer book recommendation services to different stakeholders within the

dead, the chances of publishers, let alone their technologies, to be considered important influences on a work's creation diminish rapidly.

¹⁹ M. Puchner, *The Written World: How Literature Shaped Civilization*, (Granta, 2017).

²⁰ L. Snel, GJ.

market.²¹ This thesis will not try to pre-emptively predict whether text mining will become a company, software, and in-house department, or even a new stakeholder within the market. As the work of Schweizer and Murray proves, such early predictions are too complex to be fruitful. This thesis will analyse the potential effects of text mining on the publishing process, were it to be made a part of it in any way, and show how the technology can thus impact fiction at content-level.

The reason text mining seems to be picked up only slowly by publishers appears to be related to the dynamic of dread at play between stakeholders. On one hand there are the trade fiction publishers, struggling corporations afraid of ending up lagging behind in terms of technologic advancements and thereby, they believe, profit.²² On the other hand there are the authors, afraid the introduction of text mining into the business models of their publishers means computers might steal their jobs.²³ In a third corner, one finds the critics, afraid of what publishers' blind trust in text mining might mean for the future of literature. The consequential triple fear of publishers, to lag behind innovatively, lose face with critics as well as the general public, and lose the faith of their employees and authors²⁴ leads either to their open rejection of, or to their secretive experimentation with text mining.

²¹ L. Snel, Interview with Bookarang founders, 13 October 2017.

²² H. Chin-A-Fo and T. Jaeger, "Van alleen maar kosten besparen, word je depressief", *NRC*, 16 June 2017 (1 July 2018).

²³ This explains the subdued success of Jockers and Archer's book. Not all authors are keen on seeing their creative work being subjected to the consultancy of a machine.

²⁴ H. Chin-A-Fo and T. Jaeger, 'De lezer heeft altijd gelijk, je moet hem volgen', 10 July 2017 (1 July 2018).

Gwena Jouen of WPG is certain that publishers are interested in what text mining can do for their business, but do not dare to show it after the negative media backlash on WPG's bestsellers experimentation.²⁵ In 2016, WPG ran an experiment in line with the work of Archer and Jockers. They did so with the help of the Dutch Digital Humanities research institute Huyghens and corpora from ten years' worth of published novels by WPG publisher De Bezige Bij.²⁶ WPG wanted to know whether Huygen's text mining engine could help predict, based on the novels' textual content, which ones belonged to the corpora of bestsellers and which ones to the worstsellers. They hoped it could help them see on what kinds of novels they were losing profit.²⁷ As both critics and academics have argued, WPG's text mining-based methodology appears to be insufficient support for their commercial question, an issue that will be returned to in chapter 1 and 2. The engine was, as WPG and the Huygens Institute proudly proclaimed, 78,3 percent successful at the task, making the experiment the incentive for WPG to start its own in-house R&D department Schwung. Schwung has by now developed 61 products of which they hope to launch some onto the book market and sell to projected stakeholders.²⁸

Although these first experimentations with text mining in publishing have been the motive for this thesis, they are not the subject of the first chapter. Chapter 1 will first discuss the origins, possibilities and limitations of text mining technology, and will

²⁵ L. Snel, GJ.

²⁶ H. Chin-A-Fo and T. Jaeger, "Van alleen maar kosten besparen, word je depressief".

²⁷ L. Snel, GJ.

²⁸ Ibid.

only them move on to explore its applications within publishing business operations in Chapter 2. Starting with the technology's possibilities, rather than with the trade fiction publisher's business needs, avoids promoting the uninformed and incorrect use of the technology which publishers like WPG might currently risk, an argument that will be further explained in Chapter 1.

Evidentially, both the phenomenon of text mining being used in publishing, and that phenomenon being a research topic are new. This makes finding useful and reliable sources on them hard. That is the price one pays for breaking fresh ground. This thesis therefore works with a mix of sources and research methodologies. First of all, various literature from the academic fields connected to text mining have been used to help comment on the workings and possible applications of text mining. Second of all, news articles and popular science books have been consulted in order to provide information on the latest uses of text mining in the publishing industry. Third of all, informational interviews with different employees at Dutch publishers and publishing concerns, as well as the written reports of stakeholder symposia, have been used to gain direct insight into the current use of text mining. This methodology of combining academic and non-academic sources proved to be best practice with regards to this emergent research topic.

This thesis works with terminology like 'text mining technology', 'data mining' and 'business operations'. These terms will be more elaborately explained in chapter 1 and 2 respectively. The term 'trade fiction publisher', however, might need elucidation up front. 'Trade fiction publisher' refers to a non-academic publisher that commercially sells, at least, fiction

books.²⁹ Although this thesis argues text mining can change literature, it does not use the term ‘literary publisher’. That is, first of all, because not only literary publishers publish what academia may eventually call Literature. The term stems from Bourdieu.³⁰ To him, ‘literary publisher’ signifies a type of publisher whose main concern is contributing to the cultural sphere through the publication of new literary works. Yet, this thesis defends how text mining can affect the fiction produced by any type of trade fiction publisher. Moreover, most modern-day publishers are a hybrid³¹ of cultural and financial concerns.³² That is why ‘trade fiction publisher’ is the more appropriate and workable term.

The uses of the terms ‘fiction’, ‘literature’ and ‘trade fiction publisher’ also require elucidation. This thesis limits itself to proving how fiction and, thereby, literature can be changed by trade fiction publisher’s text mining use.³³ ‘Literature’, refers to the extension, to the produced part of ‘fiction’. The useful definition of Claire Squires does not quite cover it. Squires

²⁹ G. Clark and A. Philips, *Inside Book Publishing*, (New York: Routledge, 2008), p. 275.

³⁰ F. de Glas, ‘Authors’ oeuvres as the backbone of publishers’ lists: Studying the literary publishing house after Bourdieu’, *Poetics*, 25.6 (1998), p. 379-380.

³¹ M. Bhaskar, *The Content Machine: Towards A Theory Of Publishing From The Printing Press To The Digital Network*, (New York: Anthem Press, 2013), p. 147.

³² There is reason to believe, as will be discussed in chapter 2, commercial, non-specialist publishers are likely to employ text mining for the generation of new genre fiction.

³³ Non-fiction and other genres will not be discussed, which is not to discourage others from doing so. Observations made during the research for this thesis simply showed that text mining is mostly associated with literature. That being said, the startup Bookarang recently shared an idea with the author for a new text mining product. They wanted to make it possible for journalists and other media specialists to find novels that adhere to certain themes, topics or political stances. Should the journalist need an example of a novel that treats, for example, economy and democracy from a right-wing perspective, Bookarang could text mine an Ayn Rand title for them. One can imagine that such a tool could also be developed around non-fiction books.

presents 'literary fiction' as a genre that is created and defined by the book market's mechanisms.³⁴ However, the term 'fiction' in this thesis does not refer to what the book market considers literature, but to what the academic field of Literary Studies would potentially describe as such. Not every work of fiction is a work of literature, but some are. This thesis will prove how text mining, by being used as a publishing aid, can affect the content of fiction. Thus, if the content of fiction is affected, the content of literature is affected as well. These definitions have been adopted with an eye on the aforementioned objectives of this thesis.

Chapter 1 will provide the necessary background on the origins, potential, but also limitations of text mining technology. This overview helps one understand what text mining is capable of when used in fiction publishing, thus avoiding uninformed exaggeration of the technology's powers. Chapter 1 will also argue how important it is for publishers wishing to work with text mining to know what exactly the technology does, where it comes from and what it is able to do with regards to research. If such understanding is lacking, as will be argued, publishers cannot have any control over text mining's power to affect their work.

Chapter 2 will take the 'lessons' from chapter 1 and use them to postulate realistic ways in which text mining could be harnessed to different business operations of a trade fiction publisher. Three case studies will be discussed which will show how acquisition, commissioning, editing, marketing, and sales would be affected by the involvement of text mining. The case

³⁴ C. Squires, *Marketing Literature: The Making of Contemporary Writing in Britain*, (Basingstoke: Palgrave Macmillan, 2009), p. 5.

studies will demonstrate how support, rather than take-over is the most likely role text mining can play in a publishing operation.

Ultimately, two of the three case studies will help show how text mining as a publishing aid can clearly affect the content of published literature.

The conclusion will serve as a recapturing of the lessons learned in all chapters, as a final defence of the bold hypothesis, and as an encouragement of more research on this new strand within Publishing Studies. Hopefully, the thesis will by then have demonstrated how technology can impact literature's content and form, thereby hopefully inspiring Literary Studies to unite with Publishing Studies, if only to together await the first text mined novel.³⁵

³⁵ In order to avoid any disappointments, and as chapter 2 will justify, the automated writing of novels will not be covered by the scope of this thesis.

Chapter 1.

Text Mining and its Introduction to Trade Fiction Publishing

This thesis discloses how a technology has ended up being applied in a way neither its inventors nor first users could have foreseen.

Unless perhaps you are Italo Calvino. In his dreamily chaotic *If On A Winter's Night a Traveller*³⁶ the Italian author writes about a 'reading machine' able to read, reproduce but also critically assess any text. One of the machine's operators asks the protagonist whether he would mind his reading of a novel to be compared to that of the machine.³⁷ The protagonist does mind.³⁸ Calvino's inquiries are appropriate here. Could a technology surpass a human's reading of a text? Could such a technological reading affect the creation of new texts? On the former users of text mining technology already rely to be true, as will be explained below. As this thesis shall show, the latter is what could happen as a result. This chapter shall explain how text mining technology works to have that power. This chapter does not offer an abridged

³⁶ *If On A Winter's Night a Traveller* was first published in 1980.

³⁷ I. Calvino and W. Weaver (ed.), *If on a Winter's Night a Traveler* (London: Vintage, 1998), p. 217.

³⁸ Calvino connects the machine to a totalitarian regime that uses its powers to exercise extreme censorship. Whether studying the development of publishing corporations employing text mining from a sociohistorical angle could benefit from Calvino's vision, will have to be researched elsewhere.

history of text mining. Rather, it provides the whats and hows necessary to understand its capabilities and limitations when used to assist trade fiction publishing.³⁹ After describing text mining's origins, its technical makeup, and why it is important for publishers especially to know these, three chief misunderstandings about the technology will be demystified. This chapter thus offers preparatory insights for publishers as well as Publishing Studies scholars wanting to understand the potential of text mining, or use it in constructive ways. It will rectify fundamental text mining misconceptions, to which publishers are prone, by discussing the technology's most important origins, characteristics and principles. Knowledge is, after all, power.

1.1. Technological determinism

Like Calvino's musings, this thesis' hypothesis expresses a technological determinist position. That position will be accounted for before moving on to explain how text mining works in the rest

³⁹ For those wanting to know more about the technical details of text mining, the following sources would provide a good start:

- D. I. Holmes and J. Kardos. 'Who Was the Author? An Introduction to Stylometry', *Chance*, 16 no. 2 (2003).
- A. Hotho, A. Nürnberger and G. Paaß, 'A brief survey of text mining', *Ldv Forum*, 20 no. 1 (2005).
- M. G. Kirschenbaum, 'The remaking of reading: Data mining and the digital humanities', *The National Science Foundation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation*, (Baltimore: MD, 2007).
- C. Koolen and A. van Cranenburgh, 'These are not the Stereotypes You are Looking For: Bias and Fairness in Authorial Gender Attribution', *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 2017.
- D. Meyer et al, 'Text Mining Infrastructure in R.'
- E. Tonkin and G. JL Tourte, *Working with Text: Tools, Techniques and Approaches for Text Mining*, (Elsevier, 2016).

of this chapter. At the most basic level, the definitions go as follows: Technological determinism argues that technology is fundamentally out of human control, decides its own trajectory of development and application, even has its own political agenda.⁴⁰ On the other end of the spectrum stand social determinists. They believe socio-economic development and human needs are what decide the use and evolution of technology. As Van der Weel points out, neither of these extreme notions offers full assistance in the analysis of technological change in society.⁴¹ However, one may agree that technology ‘is never neutral, and invariably brings effects that were not foreseen or intended by its inventor’.⁴² Cell phones and social media have not just affected, but reformed human communication. On demand television and Netflix have changed the ways people spend their free time, which primarily means less time spent reading books and more reading subtitles. And of course, tablets and e-readers are currently, albeit hesitantly, helping change reading and book buying behaviour.⁴³ There is no reason to believe text mining is exempted from eventually showing its power to affect change, especially not now it has caught the interest of commercial corporations. That is one reason for why this thesis exhibits a technological, rather than a social determinist stance.

Another reason for its technological determinism is the observation that, regardless of the potential effects of a

⁴⁰ L. Winner, ‘Do artifacts have politics?’, *Daedalus* (1980), p. 122.

⁴¹ A. van der Weel, *Changing Our Textual Minds: Towards a Digital Order of Knowledge*, (Manchester: Manchester University Press, 2011), p. 29.

⁴² A. van der Weel, *Changing Our Textual Minds: Towards a Digital Order of Knowledge*, p. 10.

⁴³ L. Snel, Interview with Athenaeum Booksellers’s director Caroline Reeders, 14 May 2018.

technology's use, regardless of whether the technology has been fully understood, humans show an innate drive to experiment with it. Using for the sake of using is a reality of our modern world. As Winner argues, 'the need to maintain crucial technological systems as smoothly working entities...have tended to eclipse other sorts of moral and political reasoning'.⁴⁴ Using technology, as Winner eludes to, has come to symbolise progress. This defence of technological determinism does not prove that text mining will definitely affect literature. Yet it does imply that those possible affects will appear if they can, thanks to humanity's compulsive technology use.

It is this compulsive use that is running the investigations of publishers into text mining. The problem is that their curiosity trumps their knowledgeability, causing publishers' direction of reasoning to run from their 'problem' to 'a technology'. For example, WPG wanted to save money. They hoped, and eventually believed, text mining tools could tell which books not to invest in. The tools, however, could not do so. Text mining was merely able to point out, with incomplete certainty, what features bestselling and worstselling books shared. Another example is Bookarang, who helps clients with their need to provide good book recommendations. Their clients believe book recommendations will help sell books. However, Bookarang's tools cannot say which books readers will want to buy. They can just point out which books are alike. The tendency of publishers to believe their problems can be solved by a tool is unfruitful. They will need to know what exactly a tool can and cannot tell them before they

⁴⁴ L. Winner, 'Do artifacts have politics?', p. 133.

employ it to solve an issue. Should they do so, they will have more power over the way text mining impacts their publishing processes, and thereby the fiction they produce.

Lastly, there is an argument with which to reject the applicability of the social determinist argument. It is not conscious human and societal need which is leading publishers to take to text mining. How can there be a conscious need for a technology's applications when they are still clouded in mystery? As demonstrated, at this point publishers are largely unaware of the technology's technical make-up, powers, applications and limitations. For how could they be, if those who created it are constantly evaluating and trailing them? The idea of a social need driving text mining's current integration does not hold. What the academic origins and consequent interpretation hurdles mean for that integration, will be returned to below.

1.2. What is text mining?

There are different ways to define text mining as a practice.

According to one definition, text mining is primarily a form of fact extraction.⁴⁵ This is done by extracting parts of a text and tagging these parts with appropriate attributes, like place and person's name, but also verb, part-of-speech and sentence. A second definition considers text mining to be mainly a form of data mining. The objective of text data mining here is to find useful patterns in data that have been reaped from a suitably pre-tagged

⁴⁵ A. Hotho, A. Nürnberger and G. Paaß, 'A brief survey of text mining', p. 4.

text. The second definition thus builds on processes executed in the first. The third definition, again, builds on the former two, defining text mining as a combination of fact extraction, data mining and statistical methods. Whichever way one looks at it, text mining offers a ground-breaking means to unearth information from old as well as new texts. No wonder the companies that publish texts for profit have taken an interest. As will be shown in chapter 2, text mining used as a publishing tool mostly adheres to the second and third definition.⁴⁶ One can probably already appreciate how combining data mined from text, with data stemming from the sales and finance departments may offer useable insights.

1.2.1. A superior form of reading

The average human mind is not able to read *If On A Winter's Night a Traveller* at the speed of light and count the number of times it features the word 'you'.⁴⁷ Text mining is.⁴⁸ That is why its practitioners consider text mining, as Calvino projected, an advanced form of reading. Not only does the tool offer a speed of reading texts unparalleled, it offers the collocation of data never collocated before. The innovativeness of the data that can be won with text mining is often considered the most important value of the technology.⁴⁹ Examples of such data are the number of times a

⁴⁶ Gwena Jouen of WPG/Schwung said she called what her job involved 'text and data mining'. In L. Snel, GJ.

⁴⁷ The narrative of *If On A Winter's Night a Traveller* is written in second person singular.

⁴⁸ F. N. Patel and N. R. Soni, 'Text mining: A Brief survey', International Journal of Advanced Computer Research 2.4 (2012), p. 243.

⁴⁹ F. N. Patel and N. R. Soni, 'Text mining: A Brief survey', p. 243.

verb is used to start a sentence, or the level of lexical variety, or its overarching themes. The emergent availability of such never-before perceived textual data is quietly revolutionising research in the humanities. It is not as if a human could never find what text mining can find, he or she would just need Hermione's time turner⁵⁰ and a cat's nine lives.

It seems the mythical reading capabilities of text mining have resulted in its practitioners now considering every text a dataset to-be. To them, a text not having been reduced to a dataset yet, means it still withholds its easily interpretable facts and figures.⁵¹ This outlook shows in the words of practitioner Kirschenbaum: 'Reading is not so much "at risk" as in the process of being remade'. 'While there will hopefully always be a place for long, leisurely hours spent reading under a tree,' he explains, 'this is not the only kind of reading that is meaningful or necessary'.⁵² Meyer et al back Kirschenbaum's faith by explaining that it has been nothing but a centuries-long shortage of the necessary means that has kept us from turning the 'default way of storing information', meaning text, into better structured formats.⁵³ 'Better' or 'useful' is thus to be interpreted as providing more instant access and means of exploitation.⁵⁴ The evident paradigm shift on what texts are, instigated by text mining possibilities,

⁵⁰ J. K. Rowling, *Harry Potter and the Prisoner of Azkaban*, (London: Bloomsbury Publishing, 1999).

⁵¹ The possibility that literary texts may not be able to yield 'information' and 'facts' the way an average web page or Reddit discussion can, is a topic of particularly interesting debate not covered by this thesis.

⁵² M. G. Kirschenbaum, 'The remaking of reading: Data mining and the digital humanities'.

⁵³ D. Meyer et al, 'Text Mining Infrastructure in R.', p. 2.

⁵⁴ A. Hotho, A. Nürnberger and G. Paaß, 'A brief survey of text mining', p. 2.

seems to play a part in its aforementioned compulsive use.

1.2.2. An academic field

Practitioners Kirschenbaum and Meyer, a digital humanities and a statistics scholar, are both part of the expansive and interdisciplinary academic field of text mining. Text mining is thus a practice and an academic field. The field, building on texts as datasets, has grown between disciplines like computer science, statistics, computational linguistics and information retrieval, all of which have at some points delivered new text and data mining tools and methodologies.⁵⁵ Some of the most relevant innovations pertaining to the text mining of literature have, incidentally, been developed by stylometry, a subdiscipline of linguistics.⁵⁶ Stylometry has been in development since the 1960s, when Mosteller and Wallace used statistical methods to try uncover the authorship of a selection of eighteenth-century Federalist Papers.⁵⁷ Their work lay the foundation for discoveries like Rowling being Galbraith, and Starnone writing like Ferrante.

The previous two examples indicate, as was mentioned in the introduction, that the use of text mining is proliferating. What they additionally indicate, is that the academic practice of text mining is being drawn into a commercial context. Both the algorithms and the methodologies publishing corporations use to

⁵⁵ D. Meyer et al, 'Text Mining Infrastructure in R.' p. 1.

⁵⁶ A. Hotho, A. Nürnberger and G. Paaß, 'A brief survey of text mining', p. 1-2.

⁵⁷ D. I. Holmes and J. Kardos. 'Who Was the Author? An Introduction to Stylometry', p. 2-3.

employ them stem from the academic field of text mining. It is important to realise what the scholarly origins of the technology mean for its usability in a commercial context like that of trade fiction publishing. History shows, of course, that tools can transcend their original purpose, but not without some bumps, something the new Digital Humanities subdiscipline of Tools Criticism is currently studying.⁵⁸

There are two factors impacting the integration of text mining in publishing which stem from its academic origins. The first, is that text mining scholars are constantly trialling and evaluating their own tools⁵⁹, something on which a sub-field of Digital Humanities, Tools Criticism is keeping a close eye.⁶⁰ This means that the accuracy and trustworthiness of text mining algorithms and methodologies cannot be fully guaranteed. Publishers using them pre-emptively do so at their own risk, and at the risk of their published works.

A second factor impacting text mining integration with an academic cause, is the fact that text mining algorithms and methodologies, although still being trailed and tested, are freely available online thanks to Open Source sharing.⁶¹ That is how WPG and Bookarang came by the ingredients of their tools.⁶² These eventual tools, however, are not available Open Source.

⁵⁸ For an overview of research and topics of debate within the Digital Humanities and Tools Criticism, I would advise the following sources:

- M. K. Gold, *Debates in the digital humanities*, (University of Minnesota Press, 2012).
- Digital Humanities Quarterly.

⁵⁹ C. Koolen and A. van Cranenburgh, 'These are not the Stereotypes You are Looking For: Bias and Fairness in Authorial Gender Attribution', p. 12.

⁶⁰ A. Dorofeeva, 'Towards Digital Humanities Tool Criticism', Leiden University Repository, (2014).

⁶¹ L. Snel, GJ.

⁶² Ibid.

Gwena Jouen says that, since WPG is a corporation, they do not have the authority, the figurative stamp of expertise, that allows them to share their tools with credibility like an academic institute can.⁶³ The more plausible clarification would be that they are a concern, and therefore want to be the only ones to benefit from the products they themselves develop. Sharing them for free with rival publishers would not help their competitive edge. The Open Source availability of text mining resources thus enables the technological development of trade fiction publishers, provided these publishers possess adequate R&D and financial means. The publishers who do not, have no choice but to rely on the text mining products and services of those who do.⁶⁴ Ironically, the Open Source availability of algorithms thus creates a developmental imbalance between trade fiction publishers. Moreover, it means the preferences and methodological biases of those publishers building and selling text mining aids will affect the businesses of publishers buying these aids. How such ‘bias’ can evolve, and what algorithms have to do with that, is explained in the next section below.

1.3. How does text mining work?

Text mining is done by means of algorithms, which are sequences of mathematical instructions.⁶⁵ Algorithms can be programmed to

⁶³ Ibid.

⁶⁴ WPG wishes to both employ and sell the text mining products they develop.

⁶⁵ E. Finn, *What algorithms want: Imagination in the age of computing*, (MIT Press, 2017), p. 17.

carry out certain calculations, which can involve counting, tagging, ordering, retrieving and more. Combining and stacking multiple algorithms creates an algorithm engine. To computer scientists, algorithms and algorithmic engines ‘represent repeatable, practical solutions to problems’.⁶⁶ What is important to recognise, as Finn points out, is that, ‘a method for solving a problem inevitably involves all sorts of technical and intellectual inferences, interventions, and filters’.⁶⁷ To explain what Finn means with technical and intellectual filters, Bookarang’s business will be taken as an example.

As mentioned in the introduction, Bookarang is a Dutch start-up offering text mining-based book recommendation services.⁶⁸ They provide these services using their own algorithmic recipe which compares book texts on a range of different levels, such as writing style, themes and subjects. What kinds of tags would one need in order to compare two books on the fervently discussed and disputed concept of style? Solving that question requires what Finn calls ‘technical filters’, and demonstrates one of the basic characteristics of algorithmic engines: they are biased. They cannot help but be. Finding answers demands a methodology and a methodology is built on certain assumptions. Repeating a methodology, as one does when doing text mining, thus perpetuates these assumptions, something text mining researchers are aware of and try to counteract by doing research on them.⁶⁹

⁶⁶ E. Finn, *What algorithms want: Imagination in the age of computing*, p. 18.

⁶⁷ Ibid.

⁶⁸ L. Snel, Bookarang founders.

⁶⁹ C. Koolen and A. van Cranenburgh, ‘These are not the Stereotypes You are Looking For: Bias and Fairness in Authorial Gender Attribution’.

By ‘intellectual inferences’, Finn means for example Bookarang’s aforementioned supposition that readers want to read books that resemble books they have read before. Bookarang’s concept of content-based recommendations is, of course, a maverick and potentially fruitful formula. Giants like Amazon and the Dutch Bol.com still rely on behaviour-based recommendations, for example. Nonetheless, the idea that readers want to read books that resemble books they have read before is still an inference. That is why both Bookarang’s algorithm engines and the problem the company has defined and solved with them express a certain bias.⁷⁰

Text mining, in short, offers a revolutionary way to extract data from a work of literature that may, through appropriate analysis, be turned into new knowledge. It does so by means of algorithms which are necessarily biased towards a certain goal. Publishers have access to using academic text mining algorithms and methodologies thanks to their Open Source availability. This means more affluent publishers will have better and more personalised access to text mining aid than others. The compulsive drive of some publishers to use text mining comes with the risk of using underdeveloped and overly-biased tools for the wrong problem.

⁷⁰ In 2017, Bookarang looked for research interns to help them tackle a recurring issue. Clients interested in using their services kept asking how they could be sure their content-based advice was ‘good’, was advice that made people buy more books. The company’s founders wanted the interns to help them gather proof for the validity of their advice. Perhaps the more fruitful undertaking would have been to convince clients that people buy books that are like the ones they have read already, or to at least show how come that is what their business successfully runs on.

1.4. Deconstructing misunderstandings about text mining

Finn's observations already introduced the first misunderstanding about text mining to be discussed here, namely that it automates text-based analysis, from doing literary criticism to reading a letter from your bank. Text mining does not determine what the data it delivers help to assert. That is the task of the user of text mining. The hermeneutics are left to the humans. It is important to realise that working with text mining first and foremost requires an awareness of *how* one wishes it to make meaning. WPG chose Huyghens' engine built for asserting literary quality in order to ~~try~~ understand which kinds of books sold well or badly. The algorithms did not claim that literary quality and sales figures are two sides of the same coin. That was WPG's own, albeit Huyghen's academically guided, hypothesis, which the engine partially confirmed.⁷¹ The first misunderstanding thus points back to publishers' aforementioned faulty direction of reasoning.

WPG's project provides a case-in-point for a second misunderstanding about text mining. WPG, and many text mining users with them, seem to believe text mining can draw conclusions about things it does not know. Algorithms will swiftly and efficiently extract, sort, collocate and present data mined from texts or elsewhere, but it can only do that based on data it has been fed. Vittorio Loreto, the scientist who managed to prove Domenico Starnone's style of writing was closely related to Elena Ferrante's,

⁷¹ What 78,3% success meant exactly the PR department of WPG never managed to explain.

had no data, which is to say ‘text’, written by Starnone’s wife, for whom Ferrante’s publisher’s money transfers could also have been intended.⁷² Loreto could show an expansive map showing how the writing styles and word use from certain works of Ferrante and Starnone are a category on their own, but as long as Starnone is denying the theory, and his wife does not publish any literary texts but translations, nothing has definitively been proven. Back to WPG, who believed it had confirmed the hypothesis that the higher the literary quality of a work, the better it sells. The analysis of Huygens had not accounted, however, for data on the original marketing budgets of book from the corpora. It is reasonable to suggest that marketing has a positive effect on the sales numbers of books. That is what most publishers with marketing departments hope is true. Considering the fact that the WPG experiment’s corpora consisted of the absolute best- and the absolute worstselling titles of De Bezige Bij, a 78% success rate coming from an engine that could not account for marketing budgets sounds rather unimpressive.⁷³

Finally, text mining is no magic. It works according to the laws of nature and mathematics, just like the human mind does. Just like the human mind, text mining preforms best when primed and led into the right direction when required to solve an issue. Vittorio Loreto did not have the luxury Patrick Juola had in the discovery of J. K. Rowling. Unlike Loreto, Juola had been told a rumour about Galbraith’s true identity before starting text mining

⁷² E. Etty, ‘Ferrante, who cares?’.

⁷³ L. Snel, GJ.

for answers.⁷⁴ Having evidence that allowed for the comparing of just one author with one other author made Juola's work easier and its results more reliable. Mosteller and Wallace already experienced this in the 1960s, when their statistical methods uncovered the authorship of a selection of eighteenth-century Federalist Papers.⁷⁵ They attempted to find the one-out-of-three possible authors of twelve out of eighty-five papers. Their statistical analysis of the papers' function words made them attribute authorship to the writer Madison. This finding was in line with what historians had already been predicting.

To conclude, its power to affect literature, which is the topic of this thesis, stems from text mining technology's advanced reading capabilities. The academic interdisciplinary field that nursed it into being has given it the ability to extract, index and collate data from any text. However, text mining is not a magic spell and delivers best results when primed, given the right data and, most of all, when used as part of a well-informed hermeneutic process of discovery or hypothesis confirmation. It is important for publishers to be aware of these aspects of text mining, as it holds the power to change their published literature. What that change is can only be controlled when one knows, if only superficially, how the technology works. How that power can affect that change will be looked at in the next chapter.

⁷⁴ Juola, P., 'How a computer program helped reveal J. K. Rowling as author of *A Cuckoo's Calling*', *Scientific American*, vol. August 2013, 20 August 2013 <http://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/> (9 June 2018).

⁷⁵ D. I. Holmes and J. Kardos. 'Who Was the Author? An Introduction to Stylometry', p. 5-6.

Chapter 2.

Text Mining as a Publishing Tool: Three Case Studies

When asked whether she believed text mining could assist her, editor Lisanne Mathijssen said: ‘I do not think that [it] can help predict the success of a book as early as in the commissioning phase [...]Textual analysis by editors is so much more valuable than text mining’.⁷⁶ Commissioning editor Mathijssen from HarperCollins Holland, and editor Simon Dikker Hupkes of Atlas Contact both stress the human quality of their job.⁷⁷ They believe this aspect cannot be lived up to by a conglomeration of computer algorithms.

The fear of editors primarily seems to be a result of WPG’s recent experiment – or rather the way it was presented in the media – as well as of the small hype around books like *The Bestseller Code*. Editors tend to look at algorithms the way the current president of the United States looks at immigrants. They are dangerous and disruptive, and possibly more talented than them. Even though this thesis was not written in order to calm down editors and authors, the following chapter might still do so a little. Text mining does have the ability to help fulfil editing tasks. In fact, text mining can be rendered useful in various publishing processes. The technology cannot, as Mathijssen and many others

⁷⁶ L. Snel, Interview with Lisanne Mathijssen from HarperCollins Holland, 29 March 2018.

⁷⁷ L. Snel, SDK.

fear, entirely replace those carrying out the processes. Yet, even though text mining cannot make Mathijssen lose her job, it can disrupt her job, and with that, the fiction her publishing imprint delivers.⁷⁸

This chapter discusses how text mining can be used to aid trade fiction publishing. Specifically, it takes into consideration the processes that form a trade fiction publisher's fiction publication. Through discussion of three case studies, this chapter illustrates how text mining can support the publishing processes of acquisition, commissioning, editing, marketing and sales. As the studies will show, text mining can often best be of service when combined with other kinds of data mining. Data mined from fiction texts alone – showing never before discovered intricacies and connections – are interesting material enough for scholars. Trade fiction publishers, however, require a combination of text mined and otherwise mined data for the former to serve a purpose.⁷⁹ Ultimately, the case studies will help illustrate how text mining can strengthen present habits of reasoning and handling inherent to the publishing process, thus shaping the content of the fiction that process produces.

The case scenarios are based on realistic ways in which text mining can be involved in trade fiction publishing. 'Realistic' here implies that the scenarios are based on possibilities industry

⁷⁸ It would depend on the architecture of a trade publishing corporation, and on at what level text mining would be implemented, whether the publications of individual imprints, or those of the whole corporation would be affected.

⁷⁹ 'There's a lot of information out there', said Jos de Mul, scientific director of the research institute Philosophy of Information and Communication Technology, 'that's implicit still, which you can find with data-mining', during the *Quality Non-Fiction in the Digital Era*, Dutch Foundation for Literature 8th International Non-Fiction Conference, Amsterdam, January 2011.

professionals are interested in, provided that they are technologically achievable.⁸⁰

However, 'realistic' does not imply that the following scenarios will play out at every trade fiction publisher. As was said in the introduction, this thesis has no intention of predicting exactly how and when text mining will become part of the book business. It is not exactly clear what developments to expect from which kinds of publisher, nor who will integrate text mining when and in what way. As a way to deal with this uncertainty, this thesis speaks of 'processes' being infused with text mining, rather than departments. The presence of allocated departments in the architecture of a publisher differs across the market. Processes are more universal. That is to say, not every publisher will have a sales or a marketing department, for example, yet most publishers do what can be considered marketing and sales.

Parenthetically, some say processes like marketing cannot be considered part of the practice of publishing and that it is a separate activity. However, as Bhaskar recognises, and as Squires' research confirms, marketing has in recent decennia gained influence with regard to 'content decisions, strategies and indeed creation'.⁸¹ As the second case study will show, text mining can support the phenomenon of marketing by increasing the influence on publishers' content creation. The impact of text mining thus disputes the dated designation of marketing as not being a publishing process.

⁸⁰ As was eluded to in the introduction and further explained in chapter 1, deducting possibilities of application by starting with what publishers want, rather than what technology does, is not fruitful.

⁸¹ M. Bhaskar, *The Content Machine: Towards A Theory Of Publishing From The Printing Press To The Digital Network*, p. 36.

Before moving on to the first case study, the neglect of presenting text mining as a tool with which to support the writing process will be justified. Automated writing is indeed not presented as a case study, even though it seems both the media's⁸² and scholarship's favourite topic of speculation.⁸³ There are three reasons for this. First of all, writing automation is a related practice, but not the same practice as text mining. This thesis revolves around only the latter. Second of all, both text mining scholars and text mining users⁸⁴ say there is a long way to go development-wise before the writing of a novel can be satisfactorily automated,⁸⁵ silly results notwithstanding.⁸⁶ Automating the writing process is thus currently not a technologically achievable option.⁸⁷

Third of all, although an essential part of publishing, writing is a process mainly carried out independently by the author. It is true that experiments have been done with algorithmic co-writing of literature. Right now, however, despite what Jockers and Archer expect, authors show little enthusiasm for the prospect of sharing the job with a computer⁸⁸, with the exception of some

⁸² W. de Rek, 'Cijfers En Letteren'.

⁸³ M. Roemmele and A. Gordon, 'Linguistic Features of Helpfulness in Automated Support for Creative Writing', *Proceedings of the First Workshop on Storytelling*. 2018.

⁸⁴ L. Snel, GJ.

⁸⁵ R. Giphart, 'De toekomst', in *Geen verlangen zonder tekort: De toekomst van de Nederlandstalige roman*, Dijkgraaf, M. and W. van Gils (ed.), (Amsterdam: Stichting Literatuurprijs, 2018), p. 108-110.

⁸⁶ A. Flood, 'He began to eat Hermione's family': bot tries to write Harry Potter book – and fails in magic ways', 13 December 2017 (6 July 2018).

⁸⁷ Neither is it yet a morally approved one, say publishers, authors, scholars and critics.

⁸⁸ M. Möring, 'Schrake Cijfers', *De Groene Amsterdammer*, no. 30, 26 July 2017 (8 June 2018)., and L. Snel, SDK.

critics⁸⁹, genre fiction writers and Ronald Giphart.⁹⁰ Since algorithmic co-writing is not an option yet, abs since those responsible for the writing process currently reject the option, this thesis does not consider it a case study. This being said, text mining can still impact the content of fiction through other types of integration.

2.1. No longer Lost in Translation:

Text mining and Acquisition of Foreign Titles

2.1.1. Acquisition and commission

Commissioning is the process of finding new book titles to publish.⁹¹ Publishers also seek and publish foreign works in translation, which demands fighting for and buying translation rights.⁹² What makes the job easier, however, is that foreign manuscripts are finished products. They do not still need the guidance and thorough editorial support freshly commissioned titles require.⁹³ Those who do the acquisition of foreign titles, mostly editors, need to decide whether a title is a good choice for the imprint. Examples of important factors that influence this decision are: the suitability for the publisher's or imprint's list, the promotability of the author, return on investment prospects, unique

⁸⁹ G. Post, 'De schrijver versus het algoritme', Managementblog, *Management Book Online*, 25 July 2017 (8 July 2018).

⁹⁰ R. Giphart, 'De toekomst', p. 108-110.

⁹¹ G. Clark and A. Philips, *Inside Book Publishing*, p. 96.

⁹² Simon Dikker Hupkes estimated that around fifty percent of Atlas Contact's trade titles are translations. In L. Snel, SDK.

⁹³ L. Snel, SDK.

selling points, and the quality of content.⁹⁴ Although most editors master foreign languages, to account for a large part of those factors they rely on external parties. The editor cannot pre-emptively discern the quality of content of a foreign manuscript in a language unknown to them, nor the text's unique selling points and thus potential marketability. Only after buying the rights and getting the title translated will they know the details on why a foreign title might be selling well abroad.

2.1.2. Foreign manuscript scanner

A text mining tool could assist the foreign title evaluation process by means of a manuscript scanner. This is incidentally one of the 61 products being developed by WPG.⁹⁵ The tool would evaluate the foreign manuscript by extracting its themes, topics, style and literary quality. It would be able to provide advice on different aspects according to what the editor wants to know. Moreover, the scanner could provide ways to juxtapose the ensuing metadata of the foreign manuscript to those mined from titles the publisher already owns. This could help the editor check how the foreign title compares to what they have already, or have never, published. Rather than having to rely on the foreign publisher's assurance, the manuscript scanner could help an editor execute their own assessments. The scanner could also help save editors' time in evaluating foreign manuscripts in languages they do master.

⁹⁴ G. Clark and A. Philips, *Inside Book Publishing*, p. 101-102.

⁹⁵ L. Snel, GJ.

Buying foreign rights is often experienced as a hectic and pressured affair by editors.

2.1.3. Risks and limitations

Jouen of WPG says using the scanner may lead to killing the creativity of a publisher's or single editor's list.⁹⁶ Should foreign title commissioning become decided by the extent to which a title positively resembles previously published titles, a monotone list could be the result. That is why Jouen stresses, like Mathijssen, that the acquisition of foreign titles should stay people's work.

However, using the scanner would not be the cause of such monotony, but merely of its encouragement. Editors already assess foreign titles along the lines of what the scanner can advise them on. Yet the scanner's text mining algorithms do elaborate and refine the reliance on previous results. That is to say, rather than asking the foreign editor or native speakers whether the title is of high quality, the editor could scan the title for details on lexical variety, plot structure and sentence length.

There are more risks besides monotony encouragement in using the scanner as advisory aid. Another risk would ensue from using it and ignoring chapter 1's explanation about text mining's relationship to sources. Should the scanner assist acquisition by comparing the foreign manuscript to previously published titles, it would be important to know which titles the ensuing verdict is based on. The verdict might well change according to which titles

⁹⁶ L. Snel, GJ.

are included. The foreign manuscript might be compared to a select group of previously published titles that all resemble the foreign one in certain ways. The verdict would be that the foreign title is akin to previously published work. The editor could either interpret this as a warning of repetition, or as an all-clear on appropriateness for their imprint. What might make the foreign manuscript unique, however, this basis of comparison would not allow to become clear. As explained in chapter 1, text mining cannot help evaluate that which it has not been fed. Should a foreign manuscript be unique – at least in so far as the publisher's list goes – in terms of style, lexical variety, quality or a combination of the three, the scanner cannot help show in what way. This result could also be a useful verdict, of course, but only if the one to which it is handed interprets it correctly. This will be returned to in the next case study.

Lastly, the fact that the scanner is able to encourage reliance on former titles' successes may also affect the eventual marketing of the foreign title. Since the scanner helps to partially overcome the translation hurdle, the marketing strategy can already be designed before the foreign title has even been translated. However, just like with the monotony issue, however, pre-emptively focussed marketing of titles is not caused by text mining, but could just be encouraged by it. This brings us to the next case study.

2.2. *The Tumbling of Editing's Ivory Tower*⁹⁷: Text mining and Marketing

2.2.1. Marketing

As was eluded to in this chapter's introduction, marketing is an increasingly influential power in publishers' content decisions. Before this phenomenon emerged, most trade fiction publishers considered marketing the afterthought of creation. Editors produced the title, and a copy-editor did the final checks. By then, the manuscript would be flung over the metaphorical fence of marketing.⁹⁸ Finding an audience for a new book, getting its existence into the collective consciousness⁹⁹, was not considered a part of the process of creation. That is changing.

Text mining is not the cause of marketing's increasing influence on editing. Yet, considering what the technology can help do and discover, it does have the ability to encourage and strengthen that connection..

2.2.2. Text and data mining for clues

Text mined data, especially in combination with other kinds of data, could increase marketing's influence in a variety of ways. By

⁹⁷ An expression inspired by Simon Dikker Hupkes' words, from L. Snel, SDK.

⁹⁸ L. Snel, GJ.

⁹⁹ M. Bhaskar, *The Content Machine: Towards A Theory Of Publishing From The Printing Press To The Digital Network*, p. 13.

helping to uncover patterns, they can assist in formulating arguments that support both editing and marketing decisions.

A first example of how text mining can help do that, was suggested at the end of the previous case study. Marketing can pre-emptively develop a promotional campaign based on how a new manuscript's content, style, plot and other features compare to that of previously promoted titles. Should the manuscript's content match a title whose promotion led to a big number of sales, its marketing campaign may come to resemble that of the successful title. This scenario would mostly apply to translated titles, as commissioned titles are rarely pre-emptively text-mineable. That is to say, they still need to be written when its author has signed a contract with the publisher. There would be no finished manuscript to be mined and to be fairly compared with published ones. This use of text mining, therefore, encourages the already existent methodology in marketing to promote new books the same way old books that resemble them were promoted.¹⁰⁰

A second application could also enhance present habits. Before buying a title's translation rights, or signing a contract with an author, editors may consult marketing.¹⁰¹ They want to know whether marketing considers the title promotable and sellable. With the backing-up of text and data mining, marketing has firmer ground on which either to promote a title, or reject it. This approach to text mining recalls WPG's bestsellers experiment. As

¹⁰⁰ Simon Dikker Hupkes of Atlas Contact explained how publishers will often advertise new books in folders intended for booksellers by associating them with other successful titles. The advertisement might say: 'for the fans of...' followed by a known author or title. A text mining scanner cannot, thus, be called the inventor of this custom.

¹⁰¹ L. Snel, SDK.

chapter 1 explained, should marketing decide which books to reject based on previous failures, they would do well in accounting for factors like marketing budgets in their argumentation.¹⁰²

A third application is potentially the most invasive with regard to marketing's ensuing impact on editing. Marketing could use the combination of text mined data on content, and data on sales to guide editing. Should numerous books with a certain plot structure have sold well for example, marketing might tell editing to encourage that structure in a new manuscript. Especially since author contracts get signed before manuscripts get finished, such an approach could have an impact. This application of text mining thereby has the effect of forming and perpetuating genres. Books adhering to genres like thriller, chick lit or historical fiction could be text and data mined for successful structures. The results could be used to steer the editing of new manuscripts that adhere to the same genre.

On the other hand, this text mining application might also contribute to the creation of new literary fiction.¹⁰³ This suggestion negates what many scholars believe with regards to text mining's influence on literature, which is that it can only help perpetuate existent genres. Yet, if text mining can discover and recognise genres, it can also help uncover their absence. Literary fiction does not, like genre fiction, rely on specific successful structures, but rather on unique ones. Imagine a freshly commissioned, almost-

¹⁰² Gwena Jouen explained that publishers interested in developing or using text mining tools could consider buying useful datasets from data companies. In L. Snel, GJ.

¹⁰³ 'Literary fiction' does here mean to recall Squires' definition. Squires uses the term to signify fiction that the book industry considers and sells as fiction that has literary qualities. In C. Squires, *Marketing Literature: The Making of Contemporary Writing in Britain*, p. 5.

finished manuscript which text mining shows is unlike any other manuscript published before. Marketing may want to start building a campaign for a revolutionary new work of literary fiction.

2.2.3. Risks and limitations

The use of text mining discussed in the former paragraph is what most critics, authors and editors dislike and reject. They fear it will not just lead to monotony, but to a killing of creativity. Perhaps it is some comfort to them that, judging from the meagre success of WPG's experiment, using text and data mining to find and formulate marketing and content plans barely seems helpful. Their engine was 78% successful at knowing the bestsellers from the worstsellers from corpora that contained only the absolute bestsellers and the absolute worstsellers. All in all, should marketing want to try out any of the above scenarios, they would need to know how to bring chapter 1's advice into practice. Knowing what to ask which algorithms, when to use which sources, and how to interpret the results, is essential to providing editing with scientifically sound advice.

2.3. *Strengthening Metadata: Text Mining and Sales*

2.3.1. Sales and metadata

Not every trade fiction publisher has a sales department, but all publishers sell books. One of the most important sales tools for publishers is the title information they provide retailers and other external parties with.¹⁰⁴ These metadata gets attached to the book's ISBN number – which itself is part of the metadata – and includes its title, number of pages, date of publication, but also its associated categories, themes and synopsis.¹⁰⁵ In recent years, and with the advent of new digital technologies, metadata has become a buzzword in the publishing world. Research shows that especially in the case of fiction titles, more elaborate and enhanced metadata have a positive effect on sales.¹⁰⁶ Enhanced metadata comprise the information that most influences people's decisions to buy a book, like reviews, author biographies and excerpts.¹⁰⁷ Peter Colridge, a British publishing entrepreneur, paraphrased the words of Sara Loyd, Chair of the UK Publishing Association's Digital Director's group, and told a room full of publicists that metadata is the 'most passionate bookseller'.¹⁰⁸

Metadata do not only allow retailers and direct consumers, both online and offline, to decide on buying. They also helps the publisher, or their retailer, to formulate target audiences. The more metadata a title has, the clearer it becomes who would like to read it. Knowing and acting on this, increases the chances of selling.¹⁰⁹

¹⁰⁴ F. Schut, 'Fluide metadata', *Boekblad*, 19 June 2017 (6 July 2018).

¹⁰⁵ A. Breedt and D. Walter. 'White Paper: The link between metadata and sales', *Nielsen Book Services* (2012).

¹⁰⁶ A. Breedt and D. Walter, p. 10.

¹⁰⁷ Futur Dialog, 'Metadata In The Publishing World', *Futur Dialog.com*, 12 April 2017 (6 July 2018).

¹⁰⁸ *Quality Non-Fiction in the Digital Era*, Amsterdam, January 2011.

¹⁰⁸ L. Snel, GJ and L. Snel, LM.

¹⁰⁹ One could suggest that what Bookarang does in a technical sense, is what consumers and retailers do in a literal sense: scanning and comparing metadata so as to decide on a recommendation or sale.

No wonder Eppo van Nispen, former director of the CPNB (in

Dutch: Collectieve

Propaganda van het Nederlandse Boek), said: 'I would push for more nerds in publishing companies, to understand that metadata will be the ultimate resolution'.¹¹⁰

2.3.2. Text mining for metadata

Both data expert Jouen and novice Mathijssen suspect text mining can be a vital tool in the provision of such beneficial metadata.¹¹¹

Text mining can help extract more and more intricate metadata and enhanced metadata on titles which could help PR and sales to find target audiences for them more quickly and efficiently.¹¹²

Evidentially, extra metadata could also encourage sales through retailers, like online stores and bookstores. Booksellers would have more data to base their book retailing and selling on,¹¹³ and online buyers would profit of that same increased discoverability online.¹¹⁴

2.3.3. Risk and Limitations

¹¹⁰ *Quality Non-Fiction in the Digital Era*, Amsterdam, January 2011.

¹¹¹ L. Snel, GJ and L. Snel, LM.

¹¹² G. Clark and A. Philips, *Inside Book Publishing*, p. 198-190.

¹¹³ G. Clark and A. Philips, *Inside Book Publishing*, p. 190.

¹¹⁴ H. Johnson, 'How to Sell More Books with Metadata', *Publishing Perspectives*, 27 January 2011 (6 July 2018).

Using text mining in metadata provision appears to involve only few risks. Enabling publishers and especially retailers and booksellers to rely on metadata merely risks making bookselling more superficial. Rather than having to read or scan a book, booksellers can check its metadata for pointers on what it is about and whom to sell it to. However, as was the case in the previous two studies, text mining would not create that effect directly, but merely encourage it. Faced with staggering numbers of titles to sell, retailers and booksellers are already forced to rely on publishers' metadata provisions. Elaboration of metadata, enabled by text mining, is therefore just as likely to decrease as to increase the superficiality of sales argumentation, depending on how retailers would act on the information delivery.¹¹⁵

Not a risk, but an a question that arises from publishers' enhanced metadata delivery power, is whether it could be used to create a new revenue stream. Metadata are considered important tools for selling books. They are also labour-intensive products, one that a publisher generates and consequently sees being used for free by third parties. Publishers 'used to know exactly who was using their metadata, but that is no longer the case'.¹¹⁶ Since retailers and bookseller benefit from increased metadata provision by publishers, its free supply becomes less and less straight

¹¹⁵ In 2017, Dutch chain bookseller AKO used Bookarang's services to provide visitors with a book recommendation computer. The potential buyer could tell the computer something about themselves and about their current interests. The system would use this information to recommend the potential buyer a selection of AKO books. As Gwena Jouen said about a similar WPG product, such a service saves booksellers the trouble of having to know all the books they sell. What the point would be of buying a novel from someone who does not care to read or know them, neither Bookarang nor Jouen commented on.

¹¹⁶ H. Johnson, 'How to Sell More Books with Metadata'.

forward.¹¹⁷ The use of metadata in sales does not impact the content of fiction directly, as is the topic of this thesis. It does, however, raise the need for both academics and publishers to reconsider the perspective on metadata's benefits. Text mining's influence on the shifting of that perspective forms an interesting topic of research.

Mathijssen was both right and wrong. She was right in thinking algorithms cannot predict bestsellers. Indeed their only way to help processes like editing and marketing create success, is by exposing previous success that might be emulated.¹¹⁸ However, Mathijssen was wrong in thinking text mining could never be superior to her reading. Text mining can surpass her reading in speed and pattern recognition, and thereby not replace her, but help her to rely more confidently on the coping mechanisms she employs. This also counts for her colleagues. Should her employer, the trade fiction publishing corporation, choose to take on text mining, the acquisition, editing, marketing, and sales operations could be effected. Text mining could help advance the way title acquisition and commissioning rely on previous results and marketing strategies. Moreover, it could support marketing in its advisory function to editing by enabling grand-scale and in-depth juxtaposing of title metadata and data mined sales figures. Furthermore, sales could expand their metadata provision with more and more detailed text-mined metadata on titles so as to

¹¹⁷ Bookarang says they get sent the texts of novels for free by most Dutch publishers. This allows them to extract the metadata they build their products on free of charge.

¹¹⁸ M. Möring, 'Schrle Cijfers'.

better target and increase sales. Thus, text mining makes publishing rely more heavily on methodologies already inherent to its processes.

By enabling reliance on previous results, text mining risks monotony and a rejection of creativity to prevail in published fiction. Sales might increase, but genres might be fortified and genre-defying titles never published. Affecting the publishing process means affecting what this process creates. Only the trade fiction publisher who knows exactly when to use and when to ignore text mining, may be able to avoid a potential detrimental effect.¹¹⁹

¹¹⁹ 'Yes, publishing's an instinctive business,' said Peter Colridge, 'it's an intuition-led business, but you can inform yourself with the data as well', during the conference on *Quality Non-Fiction in the Digital Era*, Amsterdam, January 2011.

Conclusion

Keeping up with how technology affects publishing does not just seem a Sisyphus ordeal. It is one. Keeping track of the latest inventions, ascertaining their capabilities, and trying to predict their uptake and effects is an endless cycle. Analysing technological innovation in a business like trade fiction publishing, one that is scorned for its conservatism, whilst being told off for progressing, is a true challenge. Those taking it on must accept that, as Bruno Latour puts it: ‘the inquirer is always one reflexive loop behind those they study’.¹²⁰

Continually pushing the boulder is still a more useful activity than ignoring or worshipping it. Rather than joining the stakeholders that see text mining spelling out the end of literature, or standing with those hailing it as the saviour of book sales, Publishing Studies can bring well-grounded nuance. The academic analysis of Publishing Studies can, and must, create some space between the polarised parties, and provide some necessary middle-ground.

Such middle-ground could serve as the fertile territory where Publishing Studies and Literary Studies end up uniting. Unlike the trade fiction publishers that reject text mining out of fear, most literature scholars reject it out of obliviousness. They hold no radical position in the debate. They have not yet joined it. Hopefully, this dissertation, as well as those that will follow it up,

¹²⁰ M. Bhaskar, M., *The Content Machine: Towards A Theory Of Publishing From The Printing Press To The Digital Network*, p. 10.

are able to convince literature scholars to join the ranks of Puchner.¹²¹ Should they let the interaction between publishing process and literary text be part of their analyses, literature scholars might be able to extrapolate intricate new levels of authorship from future, text mining-aided novels.

Text mining is many things, but not the latest job-stealing technology. Text mining is a superior form of reading in terms of pattern recognition and comparison possibilities. In the face of reading literature, it can serve as timesaver, fact checker, and inspiring third reader. Text mining provides the means with which to turn any text into an innovative dataset. Text mining is an academic practice, and now also a commercial tool. Yet, rather than a tool, text mining can best be typified as a publishing aid.

Aids cannot take your job, only assist it. They can, however thereby change it. Within the context of trade fiction publishing, text mining can assist the publishing processes of acquisition, commissioning, marketing, editing and sales. The technology's aid would strengthen existent tendencies common to those processes. The primal tendency text mining encourages, especially when combined with data mining, is: relying on tested and proven results. Text mining could help foreign manuscript assessment during acquisition, expedite marketing's increasing influence on editing choices, help create and perpetuate genres, and increase metadata resources for sales purposes. Barring the last application, these involvements of text mining in trade fiction publishing have a profound impact on the content of the published material.

¹²¹ M. Puchner, *The Written World: How Literature Shaped Civilization*.

Text mining is not a wizard, but rather its wand. It functions as well, as its user employs it. Text mining does not seek answers out of its own accord. It performs best when fed with the right data with which to answer the right question. It is the users who need to know the technology's limitations, biases and abilities in order to employ it as part of a fruitful hermeneutic process. Those who created text mining, the myriad of scholars from far and wide across the interdisciplinary field, know most of these things. Yet, these scholars have their own Sisyphus boulder, which is the continuously trailing and testing of their text mining tools. They may present their algorithms and methodologies Open Source and free of charge online, that does not mean that they are faultless. Publishers interested in using text mining should be aware of that.

Trade fiction publishers who wish to work with text mining, especially those with few technical and financial resources, may risk using it pre-emptively and gain unfruitful results. Although it is not the task of Publishing Studies to protect publishers, providing insight into new technology's possible and projected implementation is. Fulfilling this task needs more research than what this thesis has been able to cover. This thesis has, however, provided useful places to start.

First of all, more research ought to be done into how text mining could be and is being implemented into the book market. Knowing more about how and by whom the technology will be made part of the publishing processes, creates more leverage in analysing its effects on fiction. The research in this thesis indicates

general trade publishing corporations and tech start-ups may take the developmental lead as service providers.

Second of all, an accessible anthology needs to be made of the text mining algorithms and engines currently in existence and used by publishers. Publishers may not give away the recipes behind their text mining tools, but their Open Source ingredients could be extrapolated. Tools Criticism could perhaps take this mission upon themselves.

Lastly, text mining's integration into publishing asks for more philosophical inquiry into the nature of societal and technological development. Is text mining a black sheep, perhaps? Text mining's impact could be traced back from work of fiction to the processes that created it, but only by association. As said, text mining does not cause new tendencies, but only strengthens new ones. Winner suggested that perhaps not the technology, but our modern belief in its power to solely bring progress is what eventually changes fiction. The phenomenon of publishers taking to text mining thus productively invigorates the technological versus social determinism debate.

Bibliography

Primary sources

Quality Non-Fiction in the Digital Era, Dutch Foundation for Literature
8th International Non-Fiction Conference, Amsterdam, January
2011.

Snel, L., Interview with Bookarang founders, 13 October 2017.

Snel, L., Interview with Simon Dikker Hupkes at Atlas Contact, 15
March 2018.

Snel, L., Interview with Lisanne Mathijssen from HarperCollins
Holland, 29 March 2018.

Snel, L., Interview with Gwena Jouen of WPG/Schwung, 13 April
2018.

Snel, L., Interview with Athenaeum Booksellers's director Caroline
Reeders, 14 May 2018.

Secondary sources

Breedt, A. and D. Walter. 'White Paper: The link between metadata and
sales', *Nielsen Book Services* (2012).

Digital Humanities Quarterly, <http://www.digitalhumanities.org/dhq>.

De Glas, F., 'Authors' oeuvres as the backbone of publishers' lists:
Studying the literary publishing house after Bourdieu', *Poetics*,
25.6 (1998).

Dorofeeva, A., 'Towards Digital Humanities Tool Criticism', Leiden
University Repository, (2014).

- Holmes, D. I. and J. Kardos. 'Who Was the Author? An Introduction to Stylometry', *Chance*, 16 no. 2 (2003).
- Hotho, A., A. Nürnberger and G. Paaß, 'A brief survey of text mining', *Ldv Forum*, 20 no. 1 (2005).
- Juola, P., 'Industrial uses for authorship analysis', *Mathematics and computers in sciences and industry*, INASE (2015).
- Kirschenbaum, M. G., 'The remaking of reading: Data mining and the digital humanities', *The National Science Foundation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation*, (Baltimore: MD, 2007).
- Koolen, C. and A. van Cranenburgh, 'These are not the Stereotypes You are Looking For: Bias and Fairness in Authorial Gender Attribution', *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 2017.
- Meyer, D. et al, 'Text Mining Infrastructure in R.' *Journal of Statistical Software*, 25 no. 5 (2008).
- Murray, S., 'Charting the Digital Literary Sphere', *Contemporary Literature*, 56 no. 2 (2015).
- Patel, F. N. and N. R. Soni, 'Text mining: A brief survey', *International Journal of Advanced Computer Research*, 2 no. 4 (2012).
- Roemmele, M. and A. Gordon, 'Linguistic Features of Helpfulness in Automated Support for Creative Writing', *Proceedings of the First Workshop on Storytelling*. 2018.
- Schweizer, T. S., 'Multimedia Giants, Literary Publishers and New Technologies: Can Culture and Business Benefit from the Change of Rules in the Book World?', *International Journal of Arts Management*, 3 no. 3 (spring 2001).

Schweizer, T. S., 'Managing interactions between technological and stylistic innovation in the media industries', *Technology Analysis & Strategic Management*, 15 no. 1 (2002).

Winner, L., 'Do Artifacts Have Politics?', *Daedalus*, 109. no. 1 (winter 1980).

News media

Chin-A-Fo, H., T. Jaeger and one correspondent, 'Big data en algoritmes moeten worstsellers gaan voorkomen', *NRC*, 23 March 2017 <https://www.nrc.nl/nieuws/2017/03/23/met-enige-aarzeling-omarmen-de-uitgevers-big-data-als-hun-redding-7528498-a1551555> (1 July 2018).

Chin-A-Fo, H. and T. Jaeger, "'Van alleen maar kosten besparen, word je depressief'", *NRC*, 16 June 2017 <https://www.nrc.nl/nieuws/2017/06/16/van-alleen-maar-kosten-besparen-word-je-depressief-11125482-a1563416> (1 July 2018).

Chin-A-Fo, H. and T. Jaeger, 'De lezer heeft altijd gelijk, je moet hem volgen', 10 July 2017 <https://www.nrc.nl/nieuws/2017/07/10/computers-gaan-geen-boeken-schrijven-11879665-a1566188> (1 July 2018).

Etty, E., 'Ferrante, who cares?', *De Groene Amsterdammer*, no. 39, 27 September 2017 <https://www.groene.nl/artikel/ferrante-who-cares> (9 June 2018).

Flood, A., 'He began to eat Hermione's family': bot tries to write Harry Potter book – and fails in magic ways', 13 December 2017 <https://www.theguardian.com/books/booksblog/2017/dec/13/harry-potter-botnik-jk-rowling> (6 July 2018).

Futur Dialog, 'Metadata In The Publishing World', 12 April 2017 <http://futurdialog.com/index.php/2017/04/12/metadata-in-the-publishing-world/> (6 July 2018).

- Johnson, H., 'How to Sell More Books with Metadata', *Publishing Perspectives*, 27 January 2011
<http://publishingperspectives.com/2011/01/sell-more-books-with-metadata/> (6 July 2018).
- Juola, P., 'How a computer program helped reveal J. K. Rowling as author of *A Cuckoo's Calling*', *Scientific American*, vol. August 2013, 20 August 2013
<http://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/> (9 June 2018).
- Möring, M., 'Schrale Cijfers', *De Groene Amsterdammer*, no. 30, 26 July 2017 www.groene.nl/artikel/schrale-cijfers (8 June 2018).
- Post, G., 'De schrijver versus het algoritme', *Managementblog, Management Book Online*, 25 July 2017
<https://www.managementboek.nl/boekblog/actueel/6591/de-schrijver-versus-het-algoritme> (8 July 2018).
- Preston, A., 'How Real Books Have Trumped Ebooks', *The Guardian*, Guardian News and Media, 14 May 2017
www.theguardian.com/books/2017/may/14/how-real-books-trumped-ebooks-publishing-revival (8 June 2018).
- De Rek, W., 'Cijfers En Letteren', *Sir Edmund*, no. 38, 23 December 2017.
- Schut, F., 'Fluïde metadata', *Boekblad*, 19 June 2017
<https://boekblad.nl/Blogs/Item/flu-de-metadata> (6 July 2018).

Books

- Archer, J. and M. L. Jockers, *The Bestseller Code: Anatomy of the Blockbuster Novel*, (St. Martin's Press, 2016).

- Bhaskar, M., *The Content Machine: Towards A Theory Of Publishing From The Printing Press To The Digital Network*, (New York: Anthem Press, 2013).
- Calvino, I. and W. Weaver (ed.), *If on a Winter's Night a Traveler* (London: Vintage, 1998).
- Clark, G. and A. Philips, *Inside Book Publishing*, (New York: Routledge, 2008).
- Finn, E., *What algorithms want: Imagination in the age of computing*, (MIT Press, 2017).
- Gold, M. K., *Debates in the digital humanities*, (University of Minnesota Press, 2012).
- Dijkgraaf, M. and W. van Gils (ed.), *Geen verlangen zonder tekort: De toekomst van de Nederlandstalige roman* (Amsterdam: Stichting Literatuurprijs, 2018).
- Puchner, M., *The Written World: How Literature Shaped Civilization*, (Granta, 2017).
- Rowling, J. K., *Harry Potter and the Prisoner of Azkaban*, (London: Bloomsbury Publishing, 1999).
- Squires, C., *Marketing Literature: The Making of Contemporary Writing in Britain*, (Basingstoke: Palgrave Macmillan, 2009).
- Tonkin, E. and G. JL Tourte, *Working with Text: Tools, Techniques and Approaches for Text Mining*, (Elsevier, 2016).
- Van der Weel, A., *Changing Our Textual Minds: Towards a Digital Order of Knowledge*, (Manchester: Manchester University Press, 2011).

