Data Visualization using Principal Component Analysis within the Digital Humanities

Kirsten Leufkens

February, 2020

Utrecht University

Master Thesis

New Media & Digital Culture

Kirsten Leufkens
February, 2020

Supervisor: Dr. Mirko Schäfer
Reviewers: Dr. Mirko Schäfer and Dr. Imar de Vries

Utrecht University
Master Thesis
New Media & Digital Culture
Department of Media and Culture Studies

Abstract

The present study investigates the application of principal component analysis (PCA) and the resulting visualizations within the scholarly field of the Digital Humanities. Recent developments in digital technologies, tools and methods have created unprecedented opportunities for the humanities. On the other hand, the increasing use of digital methods and tools calls for a critical understanding and exploration of the ways (visual) information is processed en produced. The current debate on data analysis and the use of digital methods within the Digital Humanities has catalyzed this call. This thesis will therefore investigate the application of principal component analysis and resulting visualizations within the scholarly field of the Digital Humanities. In this study the methodological aspects of processing and visualizing data using principal component analysis in 46 Digital Humanities studies are analyzed in a structural manner, resulted in concerns on the lack of transparency in the implementation of the method and the perils of objectivity in humanities research. My research found that most researchers seem to have an incomplete, reticent and closed position with respect to the method. A more transparent and reliable approach is therefore also proposed. In conclusion I argue for a more critical and informed debate on digital methods in the humanities in general.

Keywords: Digital Humanities; Data Visualization; Principal Component Analysis; Digital Methods

Acknowledgement

After an inspiring and challenging study path – Photography, Interdisciplinary Social Sciences and New Media & Digital Culture – I am proud to conclude my studies with this master thesis. I would like to use this opportunity to thank all my fellow students of the past years for their motivation, inspiration and their enjoyable contribution to this time. In addition, I would like to thank all my teachers of the past years for everything I have learned. Without them, I wouldn't have been able to develop the way I have done.

During my master's I became very keen on the possibilities and critical view on data analysis and data visualization. It has strongly motivated me to continue on this path in my future carrier. I would especially like to thank all my colleagues and supervisor, Karin van Es, of the *Utrecht Data School* for an insightful and inspiring internship. Above all I would like to thank Mirko Schäfer for all his insights, feedback and guidance during the writing of my thesis.

Furthermore, I would like to thank my family, and my father Bert Leufkens in particular, for their encouraging motivation throughout my studies.

Special thanks goes out to Joris Veerbeek, for his loving support, ideas and advice. Our inspiring discussions at the foot of Lake Garda about data criticism, Digital Humanities and data visualizations have been very precious to me.

Contents

Introduction

1

With the increased accessibility of digital methods, archives and computational power, a growing number of scholars within the humanities have started to use computational methods to explore, analyze and visualize their objects of study. Using and combining methods from the humanities, natural language processing and computer science, this rather diverse body of work is conventionally described using the umbrella term *Digital Humanities*.[1] The use of digital tools, advocates claim, produces different kinds of knowledge[2] – arguments frequently accompanied by metaphors suggesting some kind of distance, e.g. 'distant reading',[3] 'distant viewing',[4] or 'distant horizons'.[5] Larger data sets can be processed, analyzed and aggregated using (complex) computational models, allowing us to see, from a distance, patterns and trends that would otherwise not be discernible to the naked eye. However, the use of digital technologies within the humanities isn't devoid of criticism either. Using digital tools almost inevitably means reducing complex artifices to countable objects, leading many to claim a fundamental conflict with traditional research approaches in the humanities, more concerned with hermeneutical interpretation and close reading,[6] consequently representing the digital humanities as 'uncritical positivism' or 'neo-positivism'.[7] Other critiques have concentrated on the objectivity of the computer, and on various epistemological and ontological concerns, e.g. the reduction of data in the process of analysis, the interpretation of visualizations, the possibility of manipulating data and the limited technological transparency.[8]

Recently, the article "The computational case against computational literary studies" written by Nan Z. Da and published in *Critical Inquiry* has provoked many reactions within the critical debate on digital methods in the humanities.[9] According to Da, computational and quantitative research in the field of humanities lacks statistical accuracy. Focusing on

1 David M Berry, "Introduction: Understanding the digital humanities," in *Understanding digital humanities* (Springer, 2012), 1–20.

2 Francisco J Garcia Penalvo, "Digital Humanities Data Processing," 2016,

3 Franco Moretti, *Distant reading* (Verso Books, 2013).

4 Taylor Arnold and Lauren Tilton, "Distant viewing: analyzing large visual corpora," *Digital Scholarship in the Humanities* 34, no. Supplement_1 (2019): i3–i16.

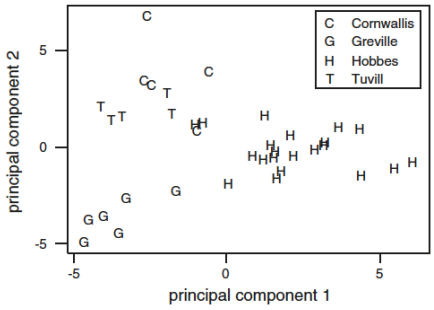
5 Ted Underwood, *Distant horizons: digital evidence and literary change* (University of Chicago Press, 2019).

6 Trevor Owens, "Defining data for humanists: Text, artifact, information or evidence," *Journal of Digital Humanities* 1, no. 1 (2011): 6–8.

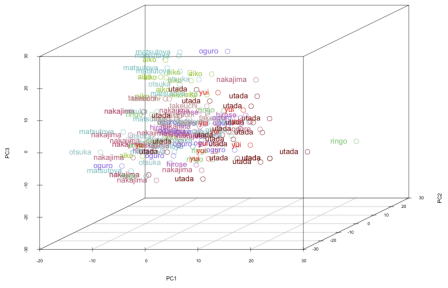
7 Tom Eyers, "The Perils of the" Digital Humanities": New Positivisms and the Fate of Literary Theory," *Postmodern Culture* 23, no. 2 (2013).

8 Cf. Bernhard Rieder and Theo Röhle, "Digital methods: From challenges to Bildung," 2017, Owens, "Defining data for humanists: Text, artifact, information or evidence"; Maureen Stone, "Challenge for the Humanities," *Working together or apart: Promoting the next generation of digital scholarship*, 2009, 43; Matthew K Gold, *Debates in the digital humanities* (U of Minnesota Press, 2012); Christof Schöch, "Big? smart? clean? messy? Data in the humanities," 2013, Christine L Borgman, "The digital future is now: A call to action for the humanities," 2010,

9 Nan Z Da, "The computational case against computational literary studies," *Critical Inquiry* 45, no. 3 (2019): 601–639.



(a) Figure adopted from: Reynolds, N. B., Schaalje, G. B., & Hilton, J. L. (2012) *Who wrote Bacon? Assessing the respective roles of Francis Bacon and his secretaries in the production of his English works.* (LLC_RSH)



(b) Figure adopted from: Suzuki, T., & Hosoya, M. (2014) *Computational Stylistic Analysis of Popular Songs of Japanese Female Singer-songwriters.* (DHQ_SH)

Fig. 1.1: Examples of various application of *principal component analyses (PCA)*

literary scholars, she argues that researchers working with computational tools are missing justification, compared to (social) sciences where large and complex data sets are reduced in a goal-orientated way. Statistical tools in the humanities, as she states, are unnecessarily and incorrectly applied in the hope that they will magically produce some interesting results by looking for patterns and answers to questions that have not even been asked yet. A technique Da discusses in her article is *principal component analysis* (PCA), which is used to reduce and simplify large complex data sets by discovering correlations in multivariate data and formulate these as principal components.[10] This technique is commonly applied within the Digital Humanities to various subjects, for example, to explore a new chronology for Shakespeare's plays or to analyze popular songs composed by Japanese female singer-songwriters (see figure: 1.1). Da calls the use of this data-reducing method an atheoretical approach to intentionally produce insightful and meaningful interpretations.[11] According to Da, statistical results are often misinterpreted, by making incorrect or implausible arguments based on the results of the analysis of e.g. word frequencies, without regard to other contextual factors. She illustrates the inaccurate application of PCA by a recreation of a PCA scatter plot of 'The Thirteen Books of Augustine's Confessions' from Piper's essay "Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel", where with a correction of the data and resizing of the distance matrices, her plot leads to different results (see figure: 1.2).

In response to Da, Piper published the article "Do we know what we are doing?" in *Journal of Cultural Analytics*, in which he compares Da's work with "the time-honoured traditions of selective reading from the field of literary criticism".[14] According to Piper, selecting a few examples to argue certain points of view is far from the same as significance testing that Da

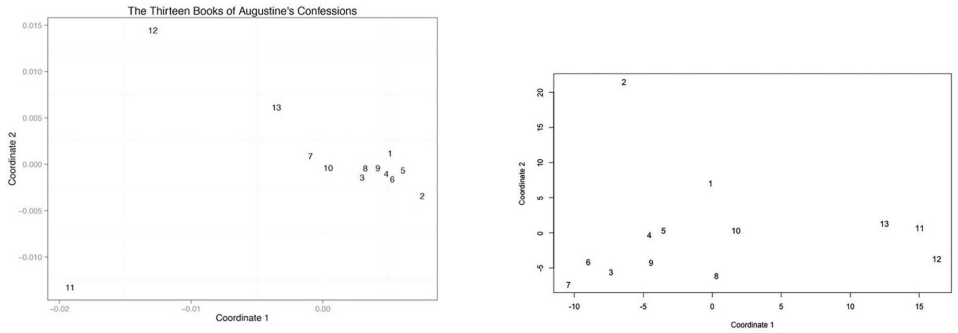
10 Schöch, "Big? smart? clean? messy? Data in the humanities"; David I Holmes, "The evolution of stylometry in humanities scholarship," *Literary and linguistic computing* 13, no. 3 (1998): 111–117.

11 Da, "The computational case against computational literary studies."

12 Andrew Piper, "There Will be Numbers Journal of Cultural Analytics," 2016, accessed September 12, 2019, `culturalanalytics.org/2016/05/there-will-be-numbers/`.

13 Nan Z Da, "The computational case against computational literary studies," *Critical Inquiry* 45, no. 3 (2019): 601–639.

14 Andrew Piper, "Do we know what we are doing?," *Journal of Cultural Analytics*, 2019, 2.



(a) "The Thirteen Books of Augustine's Confessions," from Andrew Piper[12]

(b) Da's recreated version[13]

Fig. 1.2: Illustration of inaccurate application of PCA as argued by Da

insists is necessary. Piper's main argument is the value of Da's work as proof of the need to combat the problem of generalization through the use of individual observations within scientific knowledge production. She appears to be very selective and inconsistent in her argumentation which raises doubts about the credibility of her work. According to Piper, the problem of generalizability is something that needs to be faced within the humanities, be it by reproducing work or in some other way. This thesis will contribute to what Piper stresses, as it will expand Da's selective analysis by a structural analysis of computational methods within the humanities. The conceptual and methodological shortcomings of Da's work, which Piper mentions, demonstrate the challenges faced by traditional critical models of argumentation. Renewing digital methodologies in the humanities scholars' research call for a critical understanding and investigation of the production of (visual) knowledge and the current debate on data analysis and the use of digital methods within Digital Humanities. This thesis will therefor investigate the application of principal components analysis and resulting visualizations within the scholarly field of the Digital Humanities. The research question of this thesis is: *In what way is the principal component analysis methodologically applied and visualized within the Digital Humanities?*

Previous studies have mainly presented a critical or an opportunistic perspective on digital research methods and visualization tools in the humanities through the analysis of some low sample-sized examples.[15] Criticism there was mainly focused on new aspects of the field of *Digital Humanities*, and less on how digital has changed the way research is done.[16] This research will be distinctive from other Digital Humanities reviews in its structural approach, as it will study a larger number of studies in a structural way. Because of the large number of digital methods and analysis techniques used in Digital Humanities, this thesis will focus on the processes and visualizations of a single analysis method, namely principal component

15 Cf. Da, "The computational case against computational literary studies"; Erik Malcolm Champion, "Digital humanities is text heavy, visualization light, and simulation poor," *Digital Scholarship in the Humanities* 32, no. suppl_1 (2016): i25–i32; Stéfan Sinclair et al., "Information visualization for humanities scholars," *Literary Studies in the Digital Age-An Evolving Anthology*, 2013, Helen Kennedy et al., "The work that visualisation conventions do," *Information, Communication & Society* 19, no. 6 (2016): 715–735.

16 Helle Porsdam, "Too much 'digital', too little 'humanities'? An attempt to explain why many humanities scholars are reluctant converts to digital humanities," 2011,

analysis, as discussed by Da.[17] This research will not, like Da and Piper's, complement the work that aims to introduce the idea of replication in the humanities, but will analyze the methodological issues mentioned in criticism of past Digital Humanities research. More specifically, it will analyze the methodological aspects of processing and visualizing data using principal component analysis in 46 Digital Humanities studies in a structural manner, focusing on evaluating the method section, the application of the PCA and the resulting visualizations. Despite the fact that the debate Da has called for is mainly focused on computational literary studies, this thesis will not be limited to this, but will investigate the use of PCA in the entire field of Digital Humanities. It is important to investigate the implementation of approaches and methods from other disciplines and the methodological assumptions hidden within the digital tools used in the academic field of Digital Humanities, where digital technology is bringing about a change in the way research is carried out.[18]

Particular attention in this thesis will be paid to the way in which PCA is used to visualize high-dimensional data. Due to developments in the field of visualization techniques, it is now possible to communicate ideas and analysis results better than before. The exploration and implementation of new techniques and tools for visualizing data is therefore an important research topic.[19] Within humanities, digital visualizations are becoming increasingly ubiquitous and many tools such as GIS mapping and graphs for statistical representations from other disciplines have been adopted.[20] However, people have become so familiar with these that we almost forget the meaning of these representations and how they affect knowledge production. Since visualizations seem to have an enormous convincing power, it is important that they are created and approached properly and critically.[21] In the process of creating visualizations, one has to make a lot of choices, e.g. which data sample they want to show, which indicators will support a claim the most, which visual elements they want to use. A paradigm shift seems to be taking place, which makes it increasingly important to dig deeper into the methodological assumptions that lie in the digital tools used in Digital Humanities research.[22] Enthusiasm is apparent when it comes to visualizing data, and the expectations are greater than ever before. However, it is important to analyze reflectively and critically the conditions under which knowledge is produced to avoid naive technological enthusiasm.[23]

17 Da, "The computational case against computational literary studies."
18 Rieder and Röhle, "Digital methods: From challenges to Bildung."
19 Cf. Lisa Otty et al., "Data Visualization in the Humanities," *Research Methods for Creating and Curating Data in the Digital Humanities*, 2016, Martyn Jessop, "Digital visualization as a scholarly activity," *Literary and Linguistic Computing* 23, no. 3 (2008): 281–293; Johanna Drucker, "Humanities approaches to graphical display," *Digital Humanities Quarterly* 5, no. 1 (2011): 1–21; Michael Friendly, "A brief history of data visualization," in *Handbook of data visualization* (Springer, 2008), 15–56; Min Chen et al., "Data, information, and knowledge in visualization," *IEEE Computer Graphics and Applications* 29, no. 1 (2008): 12–19; Sinclair et al., "Information visualization for humanities scholars"; Kennedy et al., "The work that visualisation conventions do."
20 Champion, "Digital humanities is text heavy, visualization light, and simulation poor."
21 Anshul Vikram Pandey et al., "The persuasive power of data visualization," *IEEE transactions on visualization and computer graphics* 20, no. 12 (2014): 2211–2220.
22 David Berry, "The computational turn: Thinking about the digital humanities," *Culture machine* 12 (2011).
23 Drucker, "Humanities approaches to graphical display."

Epistemic Criticism on Digital Humanities

2

2.1 Digital Humanities

Since the emergence of the Digital Humanities, the debate among academics about what exactly constitutes the field of research has been sparked and many papers, books and research have devoted themselves to this question.[1] It is an growing field of research that does not have a commonly shared and accepted definition yet.[2] Digital Humanities bridge a gap between the quantitative orientation of the natural sciences and the critical cultural discourse in the humanities. The use of digital methods[3] provides opportunities to generate new types of knowledge in a productive and non-traditional way, which has a significant impact on the way knowledge is generated and legitimized.[4] The use of technologies for changing the critical ground of concepts and theories within these disciplines is also referred to as the *computational turn*,[5] and goes along with an increased interest in digital and quantitative methods, included the use of computational possibilities, such as large data sets, advanced visualization techniques and processing.[6] They provide support for recognizing patterns and structures that would otherwise not be recognized.[7]

Whereas technology was originally considered to support and guide the work of humanities scholars and the emerging field of study was initially called 'computing in the humanities' or 'humanities computing', the change of the name to *Digital Humanities* was intended as a signal to indicate that it is has developed into a scholarly field – with its own research institutes, degree programs, and a growing repertoire of academic journals and books. Projects later became larger and more complex and computer techniques were developed as an intrinsic part of the research process.[8] After this shift, along with the ever-growing

1 Cf. N Katherine Hayles, "How we think: Transforming power and digital technologies," in *Understanding digital humanities* (Springer, 2012), 42–66; Edward Vanhoutte, *Defining digital humanities: a reader* (Ashgate Publishing, Ltd., 2013); Matthew G Kirschenbaum, "What is digital humanities and what's it doing in English departments?," in *Defining Digital Humanities* (Routledge, 2016), 211–220; Susan Schreibman et al., "The digital humanities and humanities computing: An introduction," *A companion to digital humanities*, 2004, xxiii–xxvii; Alan Liu, "The state of the digital humanities: A report and a critique," *Arts and Humanities in Higher Education* 11, nos. 1-2 (2012): 8–41.

2 Sander Münster and Melissa Terras, "The visual side of digital humanities: a survey on topics, researchers, and epistemic cultures," *Digital Scholarship in the Humanities*, 2019,

3 In this thesis, *digital methods* is the umbrella term for all humanities methods that make use of digital, computational and quantitative aspects. Although opinions differ, in order to maintain consistency the term *tools* refers in this thesis to the medium used within these methods, such as software applications, algorithms or programming codes.

4 Rieder and Röhle, "Digital methods: From challenges to Bildung."

5 Berry, "Introduction: Understanding the digital humanities," 11.

6 Ibid.

7 Franco Moretti, *Graphs, maps, trees: abstract models for a literary history* (Verso, 2005); Münster and Terras, "The visual side of digital humanities: a survey on topics, researchers, and epistemic cultures"; Rieder and Röhle, "Digital methods: From challenges to Bildung."

8 Berry, "Introduction: Understanding the digital humanities."

developments in digital methods, the Digital Humanities became very popular for the last decades. The change of the word 'computational' to 'digital' was not only a change of name, but the scope of the field of study broadened and expanded in its entirety.[9] This also had consequences for the object of study. Where it previously consisted of mainly literary, linguistics and stylistics studies, from then on also images, video, audio and other forms of digital media could be studied, manipulated or analysed by means of the computer.[10]

Because the use of digital tools goes beyond merely complementing traditional methods, we need to be much more aware of the impact on traditional methods, and how digital technology rather shape new methods.[11] The use of digital methods and tools is not only seen as valuable in speeding up research or enlarging the corpus. It is mainly considered to be important for raising new questions, opportunities and challenges.[12] In this way, digital influences can bring about change, insights and new ways of thinking, as a new enrichment of opportunities and possibilities. The integration and developments of digital methods in humanities scholarship has changed the way one works with their material, how they perceive and interpret it and how they work with it. Within the Digital Humanities, new methods and practices of critical reading are being developed to understand the flow of information that society produces from text, images, locations and videos. This goes beyond the point of sharing information together in a stack order to discover values and facts. Rather, these developments are important for the key idea of the university, namely *Digital Bildung*, which implies that people not only learn digital practices but think critically about the effect and process of these digital practices and culture in a computational era.[13]

Due to the fact that digital artifacts occupy everywhere in today's post-industrial society, the successes of physical machines and software are increasingly accessible to scholars. As a result of these developments, people are not only able to study non-digital objects and phenomena with the help of computers, but there also has been an enormous expansion of (digital) materials that did not yet exist or were available before the transformation of digitization and technology.[14] A distinction can be made between 'traditional' cultural artifacts that are digitally available, such as books and films, and 'born digital' cultural artifacts, for instance online publications, software programs, computer games, traces of use of online interaction, etc. More researchers are focusing on automated analysis methods to investigate these cultural artefacts.[15] Expanding on humanities research applied with digital technology, research is being conducted in the field of Digital Humanities into the impact and value of these digital technologies on the basis of humanistic insights.[16] It is

9 Edward Vanhoutte, "The gates of hell: History and definition of digital| humanities| computing," in *Defining Digital Humanities* (Routledge, 2016), 135–172.

10 Matthew Kirschenbaum, "What is digital humanities and what's it doing in English departments," *Debates in the digital humanities* 3 (2012).

11 Rieder and Röhle, "Digital methods: From challenges to Bildung."

12 Steven E Jones, *The Emergence of the Digital Humanities (Open Access)* (Routledge, 2013).

13 Berry, "The computational turn: Thinking about the digital humanities."

14 Porsdam, "Too much 'digital', too little 'humanities'? An attempt to explain why many humanities scholars are reluctant converts to digital humanities."

15 Rieder and Röhle, "Digital methods: From challenges to Bildung."

16 Chris Alen Sula, "Digital humanities and libraries: A conceptual model," *Journal of Library Administration* 53, no. 1 (2013): 10–26.

quite likely that the lack of a clear definition is also due to the broad scope of the various types of approach that research by the Digital Humanities contains.[17]

2.1.1 Data Visualization

Alongside the growing popularity of the use of digital methods for analyzing data, visualization techniques have become increasingly important in recent years.[18] More and more organizations offer new visualization tools for humanities scholars, and it is increasingly promoted at workshops and conferences. Visualization can be defined as a group of techniques and methods for creating diagrams or images to communicate a message.[19] The use of graphs as a tool for understanding information is not new – think of the prehistoric cave paintings, Egyptian hieroglyphics and Greek geometry – however, the emergence of digital technology created a lot of new opportunities and made data visualization a flourishing topic within the field of humanities.[20] New possibilities have subsequently resulted in enhanced research methods and tools, which provide new forms of visual presentation and creation of insights and knowledge.[21] Digital visualizations are becoming increasingly ubiquitous and many tools such as GIS mapping and graphs for statistical representations from other disciplines have been adopted. The power of visualization is that it makes information visible and understandable through the use of visual communication. In general, data visualizations show relations that are quantitative in nature, however, qualitative information can also be displayed, such as network analyses. Visualizing data is only effective when it translates abstract information understandable into visual representations in a way that can distinguish our eyes. This has to be done in an easy, effective and meaningful way.[22] From an epistemological point of view, visualizations create new forms of knowledge production.[23] They could be mere illustration, or translation of calculations, results, simulations into a graphic display. Data visualizations might either serve illustration of something that could as well be represented in a table or numeric fashion, but it can also be used for an analytical purpose.

However, data visualization contains such an enormous convincing power and one seems to rely more on the visual, which makes it even more important that a visualization conveys the correct information.[24] Edward Tufte has had a great deal of influence in the field of scientific data visualizations. In his books he demonstrates graphic excellence using correct and flawed examples and principles for scientific graphs.[25] The most important principles for excellence

17 Patrik Svensson, "The landscape of digital humanities," *Digital Humanities*, 2010,

18 Stefan Jänicke et al., "Visual text analysis in digital humanities," in *Computer Graphics Forum*, vol. 36, 6 (Wiley Online Library, 2017), 226–250.

19 Champion, "Digital humanities is text heavy, visualization light, and simulation poor."

20 Frank Hartmann, "Humanization of Knowledge through the Eye," *Making Things Public–Atmospheres of Democracy*, 2005, 698–707; Jessop, "Digital visualization as a scholarly activity."

21 Jessop, "Digital visualization as a scholarly activity."

22 Stephen Few and Perceptual Edge, "Data visualization: past, present, and future," *IBM Cognos Innovation Center*, 2007,

23 Rieder and Röhle, "Digital methods: From challenges to Bildung."

24 Jessop, "Digital visualization as a scholarly activity."

25 Edward R Tufte, *The visual display of quantitative information*, vol. 2 (Graphics press Cheshire, CT, 2001); Edward R Tufte et al., *Visual explanations: Images and quantities, evidence and narrative*, 1998; Edward R Tufte, "The visual display of quantitative information Graphics Press," *Cheshire, Connecticut*, 1983,

in scientific graphs that Tufte discusses are 'conciseness, clarity and accuracy'.[26] People just have become so familiar with it that we almost forget the meaning of these representations. Within the humanities there is already a group of academics actively engaged in giving a critical reflection on the use of data visualizations.[27] Stone, for example, argues the ease with which visualizations are created, and their deceptive power.[28] Criticism of visualization tools is often about the poor implementation of basic principles of visual design. Critics claim that colors used are too bold, fonts in the visualizations too small and the result would be fuzzy and misleading. Existing criticism of visualizations encompasses the idea that they favor certain points of view, strengthen existing and create new power relationships. On the other hand, the creators of visualizations believe in the stimulating power of visualizations for a better understanding of data by making data accessible and transparent.[29]

2.2 Challenges of using Digital Methods

While the possibilities of digital research methods and tools seem to be promising, there have been many criticisms of the challenges and limitations of the use of digital methods and the application of the visualization of data in the humanities. An awareness of the challenges of applications of digital methods and the epistemological and ontological effects on humanities scholars sometimes seems to be lacking.[30] In addition to the fact that digital and technological developments are changing the nature of humanities' research, there is a lack of historical background with regard to discussions on theoretical and methodological issues and underlying assumptions.[31] Notwithstanding the many ideological possibilities, a critical reflection on the use of computational tools appears to be increasingly important and necessary than ever. A critical approach and reflection to the possibilities, but maybe more the limitations of the application of computational and digital methods, led to the emergence of *Critical Digital Humanities*. The critical approach to digital and computational application – evolved from software studies and critical code studies – is introduced by Berry as the *third wave*, which "concentrated around the underlying computationality of the forms held within a computational medium [...] to look at 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".[32] This thesis will build on four major challenges of using digital methods in Digital Humanities, with regard to the visualization of data, which the critical debate is concerned with: 1) the manipulation and reduction of data, 2) the technological transparency of computational tools, 3) the seduction of scientific objectivity

26 Stone, "Challenge for the Humanities," 44.
27 Cf. Jessop, "Digital visualization as a scholarly activity"; Drucker, "Humanities approaches to graphical display"; Champion, "Digital humanities is text heavy, visualization light, and simulation poor"; Sinclair et al., "Information visualization for humanities scholars"; Otty et al., "Data Visualization in the Humanities."
28 Stone, "Challenge for the Humanities."
29 Kennedy et al., "The work that visualisation conventions do."
30 Hayles, "How we think: Transforming power and digital technologies."
31 Jessop, "Digital visualization as a scholarly activity."
32 Berry, "The computational turn: Thinking about the digital humanities," 4.

and 4) the importance of interpretation of results and visualizations.[33] The four challenges are explained in detail below.

Data Reduction

The most commonly used methods within Digital Humanities are not necessarily aimed at verifying or falsifying hypotheses, but uses digital computational techniques to generate meaningful information from a data set in order to give meaning to the data – often using impressive visualizations – by revealing new perspectives, relationships or characteristics that have not been predicted or thought through in advance.[34] Criticism about data reduction in Digital Humanities concerns the lack of purposefulness, justification and statistical accuracy.[35] Another concern within the criticism of the application of statistical calculations from e.g. information sciences, social sciences and biology, is whether these can simply be integrated into the digital environment of humanities or need to be adapted to the conventions, procedures, research questions and standards of the humanities. Johanna Drucker argues the risk for the humanities of blindly adopting methods, including the epistemological biases that are embedded in all this.[36] According to her data visualizations can sometimes turn out to be 'a kind of intellectual Trojan horse',[37] creating unseen hazards by giving 'unquestioned representations of "what is"'.[38] Above that, visualizations often have a high aesthetic value that provides a powerful argumentation of visual evidence with a convincing impact. Visualizations make certain aspects of a phenomenon visible, but only a reduced image is represented. This reduction can originate in technical aspects, but also in manipulative intentions to display or not to display certain information. Despite the reduction in information, visualizations can have an enormous persuasive impact.[39]

Transparency

In *Critical Digital Humanities: The Search for a Methodology*, a book that offers a demonstration of the critical commentary on the computational turn in the humanities, Dobson argues that "[...] computational tools are not transparent nor are they value-free".[40] He refers to the embedded assumptions and concepts in computational tools, which are not always in line with the objectives of the humanities. According to Dobson, greater methodological awareness and self-criticism is therefore required when computational tools derived from

33 Cf. Rieder and Röhle, "Digital methods: From challenges to Bildung"; Da, "The computational case against computational literary studies"; Kennedy et al., "The work that visualisation conventions do"; Johanna Drucker, *Graphesis: Visual forms of knowledge production* (Harvard University Press, 2014); Drucker, "Humanities approaches to graphical display"; James E Dobson, *Critical Digital Humanities: The Search for a Methodology* (University of Illinois Press, 2019); Lorraine J Daston and Peter Galison, *Objectivity* (Zone Books, 2007); Tufte, *The visual display of quantitative information*.

34 Rieder and Röhle, "Digital methods: From challenges to Bildung."

35 Da, "The computational case against computational literary studies."

36 Drucker, "Humanities approaches to graphical display."

37 Ibid., 1.

38 Ibid.

39 Rieder and Röhle, "Digital methods: From challenges to Bildung."

40 Dobson, *Critical Digital Humanities: The Search for a Methodology*, 6.

the empirical sciences are applied in the Digital Humanities.[41] He compares the disguised, complex and abstracted processes that hide behind the computational tools and embedded algorithms with a so-called *black box*.[42] They function as an operator – sometimes regarded as magical – where data is stored and from which meaningful and valuable results are expected, without knowing exactly what is happening or having any insight into it. The automated steps of the algorithms are often too complex to understand in detail which steps have been made and to explain which choices have been made. Not only the algorithms, but also the data structures and the underlying modeling of the digital methods are not always understood or correctly applied.[43] Dobson emphasizes the importance of reflexivity within the digital humanities in the use of computation, to prevent the misappropriation of interpretative work that is embedded in computational tools. From the moment the data is selected, filtered, composed up to the analysis, transformation, manipulation and visualization.[44]

Objectivity

Many of the criticisms have strong claims to objectivity and the misconception that automatic data processing and digital analysis would lead to less human mistakes and subjectivity, which is also called the 'lure of objectivity' by Rieder and Röhle.[45] The use of digital methods would enhance the objectivity of the research, compared to the interpretative qualitative work of the traditional humanities. The integration of digital methods in the humanities is also seen as a desire to compete with the ideal of scientific objectivity in the natural sciences through the use of machines. The book *Objectivity* of historians of science Daston and Galison provides a historical overview of the shifting ideas objectivity affected by scientific practice.[46] After the truth-to-nature principle, where botanists and naturalists collaborated with artists and painters in search of the truth and beauty of nature in a single reasoned image, the conviction arose that the production of knowledge should be disconnected from human subjectivity, and this resulted in the idea of objectivity. Developments and concerns about the human intervention in science created the idea of mechanical objectivity; the idea that production of knowledge should be free from subjectivity and that this can only be achieved through the use of machines. From 1900 on wards, however, the idea of trained judgment was introduced, in which the human interpretation of experts was central to machine processing. This also relates to the use of computers in the humanities for knowledge production, in which the computer can never be separated from subjectivity. Due to the low level of expertise of technological tools and quantitative calculations, computational analysis can still lack human subjectivity, because the expertise of trained judgment seems to be missing. Besides that, carrying out quantitative research using computational automation also goes hand in hand with subjective choices and assessment.[47] This is no different for the

41 Dobson, *Critical Digital Humanities: The Search for a Methodology*.
42 Ibid., 8.
43 Rieder and Röhle, "Digital methods: From challenges to Bildung."
44 Dobson, *Critical Digital Humanities: The Search for a Methodology*.
45 Rieder and Röhle, "Digital methods: From challenges to Bildung," 71.
46 Daston and Galison, *Objectivity*.
47 Rieder and Röhle, "Digital methods: From challenges to Bildung."

application of data visualization within the Digital Humanities. Data visualizations often have a high aesthetic value that provides a powerful argumentation of visual evidence with a convincing and persuasive impact. They may give the impression of showing the facts and reality through the intended objectivity and quality. However, a visualization is always a representation of reality. Kennedy et al. argue that the sense of objectivity is not due to the designers of the visualizations, but due to the conventions on which the visualizations are built and produced.[48]

Interpretation

In addition to the subjectivity of methodological assumptions, data are always manufactured and created by humans. The researcher is faced with many decisions during all phases of the research – choices have to be made in advance, during the production of the data set, data management and coding. In this way, data is also seen as a cultural artifact, which depends on human processing and interpretation. The production of knowledge is just as much an interpretative action in the Digital Humanities research as the coding, mining, manipulation and exploration of data.[49] The growing data stream needs new technologies to process, and thus to interpret the relevance and significance of the data. A striking feature of data visualization is that it can create new insights, and thus perhaps goes beyond just displaying the results.[50] On the one hand the visualization is an object of research that is used to investigate specific phenomena, whereas on the other hand the visualization has a communicative function to present the research results. The difference between numerical and visual reasoning compared to textual, is that the former seems to be accepted as evidence rather than argumentation since it is less likely to be questioned. It would be important to focus on the decisions and reductions of the information made during the process of producing the visualization and to critically reflect on the meaning, validity and generalizability of the visualization.[51] The idea of interpretation of quantitative data and visualizations seems to be a paradoxical issue. While Moretti is convinced that 'quantitative data are useful because they are independent of interpretation',[52] it seems more important to Digital Humanities to avoid black boxes of Trojan horses and that they should move more into the direction of Drucker's idea, who argues that we should rather consider data as capta: 'taken not given, constructed as an interpretation of the phenomenal world, not inherent in it'.[53]

These four challenges of using digital methods and tools in Digital Humanities will be operationalized for a reading of a sample of texts. In the next chapter I will elaborate on this methodologically.

---

48 Kennedy et al., "The work that visualisation conventions do"; Rosemary Lucy Hill et al., "Visualizing junk: Big data visualizations and the need for feminist data studies," *Journal of Communication Inquiry* 40, no. 4 (2016): 331–350.

49 Owens, "Defining data for humanists: Text, artifact, information or evidence."

50 Champion, "Digital humanities is text heavy, visualization light, and simulation poor."

51 Rieder and Röhle, "Digital methods: From challenges to Bildung."

52 Moretti, *Graphs, maps, trees: abstract models for a literary history*, 30.

53 Drucker, "Humanities approaches to graphical display," 1.

Methodology

3

In this chapter the methodological design of this thesis will be explained. This thesis will analyze the methodological aspects of processing and visualizing data using principal component analysis in 46 Digital Humanities studies in a structural manner, focusing on evaluating the method section, the implementation of PCA and the resulting visualizations. First of all, the data collection will be described, then the analytical rubric and additional models that has been drawn up for this structural research will be explicated and substantiated.

3.1 Data collection

The corpus for this study consist of academic Digital Humanities articles that made use of principal component analysis. The conceptual and methodological shortcomings, as lamented by Da and defensively reacted to by Piper, will be included in this study in order to counteract problems of generalization and, as Piper calls it, "poorly-sampled evidence".[1] Therefore, the corpus that will be used in this research will consist of a relative large number of academic articles, obtained from three journals; *Digital Scholarship in the Humanities*, *Digital Humanities Quarterly* and the proceedings of the *Annual Conference of the Alliance of Digital Humanities Organizations*.

The selection of the journals that will be used in this thesis is based on the corpus selection of similar research from Jänicke et al.[2] They provide a classification of close and distant reading visualization techniques, applied in Digital Humanities articles, looking for underlying characteristics and challenges compared to techniques used by information visualization experts. Although the focus of these articles is on the developments of close and distant reading, it is a similar research to this thesis since they have also analyzed a large number of papers, namely 84, of major Digital Humanities journals in a structural manner. Because of financial, political or institutional reasons, but especially because of the diversity of the work of Digital Humanities, the location of Digital Humanities works varies quite often.[3] In this thesis, a conscious choice has been made to maintain this comparable selection, and not to switch to, for example, Computer Science or Computational Linguistic journals in order to keep it comparable and falsifiable for follow-up research – and, of course, to stay within the scope of this thesis.

1 Piper, “Do we know what we are doing?,” p.11.
2 Jänicke et al., “Visual text analysis in digital humanities”; Stefan Jänicke et al., “On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges.,” in *EuroVis (STARs)* (2015), 83–103.
3 Sula, “Digital humanities and libraries: A conceptual model.”

Tab. 3.1: Overview of the selected journals and proceedings

Journal/Proceedings Papers	Papers
Proceedings of the Annual Conference of the Alliance of Digital Humanities Organizations	13
Digital Scholarship in the Humanities/ Literary and Linguistic Computing	28
Digital Humanities Quarterly	5

3.1.1 Corpus

Through the publicly available archives on the websites of *Digital Scholarship in the Humanities/Literary and Linguistic Computing*[4] and *Digital Humanities Quarterly*[5] the articles have been collected via the databases. These are both publicly available and the papers have been peer-reviewed. In addition, they are covered by various indexing services, such as *Emerging Sources Citation Index* and *Scopus*. Through the web pages I have collected the PDF files of the selected articles. For the *Annual Conference of the Alliance of Digital Humanities Organizations* this was more difficult, since there are separate websites for the different conferences per year on which the abstracts are published. For the 2014 and 2015 conferences, the abstracts could not be retrieved. The query 'principal component analysis' was used to collect the articles that met each of the following requirements:

- The research described in the article has made use of the principal component analysis
- The article contains one or more visualizations of the results of the principal component analysis
- Only research articles will be included, reviews or case studies will not be included

With the retrieved articles I built a corpus of primary sources to be analysed for this thesis. The corpus consists of a total of 46 articles obtained from the three aforementioned journals, published between 1996 and 2018 (see table 3.1). An overview of the whole corpus can be found in appendix 6.1. The articles are labelled according to the abbreviations of the journals and the initials of the corresponding authors. Figures 3.1 and 3.2 show the distribution of the number of articles over years and countries.

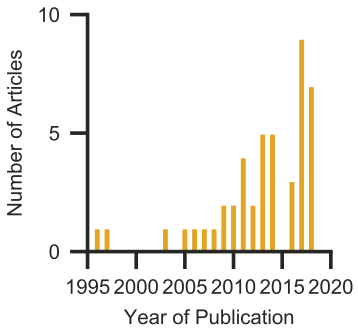


Fig. 3.1: Number of selected articles per year

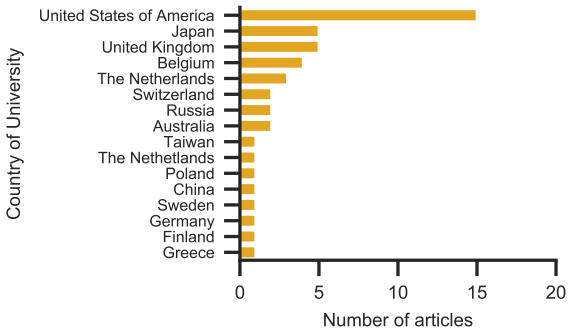


Fig. 3.2: Number of selected articles per country

[4] https://academic.oup.com/dsh
[5] http://www.digitalhumanities.org/dhq/

3.2 Analysis Procedure

3.2.1 Principal Component Analysis

The method on which this research has focused is *principal component analysis* (PCA). PCA is a statistical method that is probably the most commonly implemented in statistical software packages.[6] It is a method that is based on the idea of reducing and standardizing the dimensionality of multivariate data sets by means of a transformation to a new set of variables; principal components. This is done while maintaining as much as possible of the variance present in the data set. Instead of displaying the data set with many variables as a large mass of numbers, it reduces the data set to two or three dimensions to give a simple visual representation of multivariate data.[7] PCA is mainly used to reduce a multivariate data set and to recognize clusters and outliers within the data.[8] The visual representation of PCA results is a scatter plot, a graphical representation that shows a certain amount of the variance of the data, where the first principal component represents the major part of the variance, and the second (third, fourth, etc...) following it.[9] This relatively simple and modest way of data visualization is considered to be a diverse and useful type of statistical graphics.[10]

3.2.2 Analytical Rubric and Annotation Model

In order to investigate whether the researchers of the selected articles are aware of the challenges of using digital methods, which are central to this research, I created an analytical rubric (see table 6.2 in appendix 6.2). The rubric is based on the principles of the London Charter, which is a document with six established fundamental methodological principles for visualizing data and is created for computer-based visualization of cultural heritage.[11] Drawing from the four challenges listed above, I expanded the London charter principles with categories and associated aspects drawn up on the basis of principles and standards of the application of the principal component analysis, elements from visualization conventions and challenges identified in critical studies on the use of digital methods in the humanities.[12] The purpose of the analysis on the basis of the rubric is to gain insight into the extent to which the authors of the articles point to an awareness of these challenges, express a critical

6 Chun-houh Chen et al., *Handbook of Data Visualization* (Springer, 2008).
7 Ian Jolliffe, *Principal component analysis* (Springer, 2011).
8 Svante Wold et al., "Principal component analysis," *Chemometrics and intelligent laboratory systems* 2, nos. 1-3 (1987): 37–52.
9 Muzammil Khan and Sarwar Shah Khan, "Data and information visualization methods, and interactive mechanisms: A survey," *International Journal of Computer Applications* 34, no. 1 (2011): 1–14.
10 Michael Friendly and Daniel Denis, "The early origins and development of the scatterplot," *Journal of the History of the Behavioral Sciences* 41, no. 2 (2005): 103–130.
11 Richard Beacham et al., "An introduction to the London charter," *Joint Event CIPA/VAST/EG/EuroMed*, 2006,
12 Wold et al., "Principal component analysis"; Beacham et al., "An introduction to the London charter"; Kennedy et al., "The work that visualisation conventions do"; Da, "The computational case against computational literary studies"; Piper, "Do we know what we are doing?"; Drucker, *Graphesis: Visual forms of knowledge production*; Tufte, "The visual display of quantitative information Graphics Press"; Jolliffe, *Principal component analysis*.

attitude towards them or – at best – are reflective and transparent on the choices and actions within their research.

The focus of the rubric will be on the four challenges of using digital methods in Digital Humanities, with regard to the visualization of data, which the critical debate is concerned with: the manipulation and reduction of data, the technological transparency of computational tools, the seduction of scientific objectivity and the importance of interpretation of results and visualizations. This study aims to gain insight into the awareness, formulation and implementation of the reflexivity of the researchers in a qualitative and quantitative way.

Before analyzing the articles according to the analytical rubric, characteristics of the articles and associated visualizations were annotated in a schematic manner in order to gain insight in an even more detailed way than the rubric (see appendix 6.3). For example, the manifestation of critical reflection is charted by means of the rubric, and the second annotation model presents, among other things, the argumentation of the application of PCA. In addition, the model includes features mentioned by Drucker in her book *Graphesis: Visual forms of knowledge production*, in which she argues the importance of interpreting visual forms not just as presentations of knowledge, but as productions of it. She provides a guide to doing humanistic data visualization well, illustrated by 'windows', which she uses as a lens for interpreting data visualizations. The windows, which she argues for, the Gestalt principles and tendencies and the basic graphic variables identified by Jaques Bertin, have been used to gain insight into the way in which the results of PCA are displayed graphically.[13] The indicators on which these models are based are shown in table 3.2 and drawn up using visualization principles from Tufte,[14] Gestalt (as described by Drucker[15]), London Charter[16] and conventions proposed by Kennedy.[17]

Tab. 3.2: Overview of the Indicators

Category	**Aspects**
Aims and Methods	Method Argumentation
	Transparency of Method
	Accessibility of Data
Analysis Process	Type of Data
	Corpus Size
	Number of Dimensions
	Variance
Results	Interpretation of clusters/patterns/outliers
	Generalization
	Application of Further Examination
Visualizations	Clear Layout
	Axes and Scales
	Colours/ Shapes/ Sizes
	Titles, Labels, Legend, Caption

13 Drucker, *Graphesis: Visual forms of knowledge production*.
14 Tufte, *The visual display of quantitative information*; Tufte et al., *Visual explanations: Images and quantities, evidence and narrative*.
15 Drucker, *Graphesis: Visual forms of knowledge production*.
16 Beacham et al., "An introduction to the London charter."
17 Kennedy et al., "The work that visualisation conventions do."

Analysis and Results

4

In this chapter, the core findings of the analyses by means of my own defined analytical rubric and schematic annotation model, as described in section 3.2.2, will be discussed. The results have been divided into the four challenges as discussed earlier: *data reduction*, *transparency*, *objectivity* and *interpretation*. In the 46 articles analyzed, 141 scatter plots as a result of the application of principal component analysis – of a total of 173 scientific visualizations – were analyzed. Figure 4.1 shows how the PCA visualizations relate to the total number of scientific visualizations in the corpus and how they distribute among the different papers. Other scientific visualizations consisted mainly of results of hierarchical cluster analysis and box plots of frequency samples or distribution of topic scores. As described earlier, the subjects of Digital Humanities studies are very different and varied,[1] which is also reflected in the corpus of this study. Where the majority of studies are on subjects such as literature, for example to verify authors of books and texts, the remaining subjects range from Japanese popular music [DHQ_SH] to Dutch dialects and surnames [LLC_MHN]. Most of the studies are text-based subjects such as gospels, novels, language, music, dialogues, and poetry. Deviating and non-text-based subjects are stone artifacts [DHQ_Sci] and newspaper layouts [ADHO_BKX]. According to Da, the quantification of text, for example on the basis of word frequencies, is not transferable to the humanities discipline. This research shows that it is not applied to something complex as a genre, but rather for authors' style, something for which it has been proven that word frequency analyses can capture it well.[2]

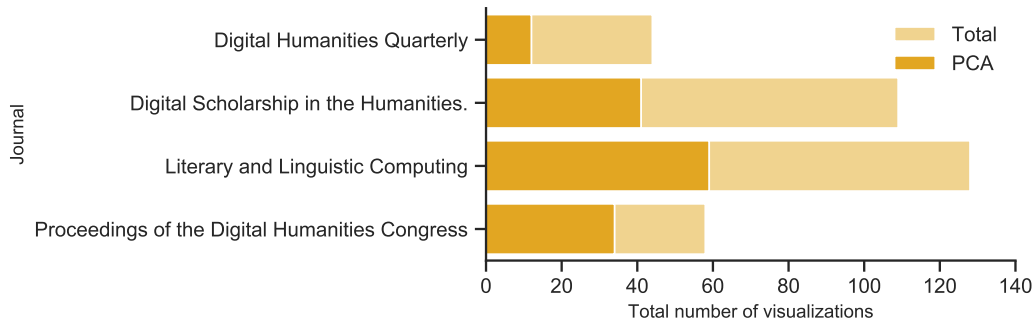


Fig. 4.1: Number of visualizations in the corpus

1 Svensson, "The landscape of digital humanities"; Vanhoutte, "The gates of hell: History and definition of digital| humanities| computing."

2 John Burrows, "'Delta': a measure of stylistic difference and a guide to likely authorship," *Literary and linguistic computing* 17, no. 3 (2002): 267–287.

4.1 Digital Humanities Challenges

Figure 4.2 shows the results of the analysis using the analytical rubric in a box plot. It gives an overview of the distribution of the results of the analysis. The results show that the authors seem to revealing an conscious and/or critical position in the articles with regard to most aspects: *Aim of method*, *process*, *limitations*, *self-reflection*, *interpretation* and *clean layout*. They seem to be more reflexive with respect to *data* and the application of *further research*. For the aspects *description of method* and the use of *scales*, the results show no or an inaccurate awareness of application. In the following sections, the results will be further discussed. The complete results can be seen in appendix 6.4. In the following sections I will discuss the results for each challenge.

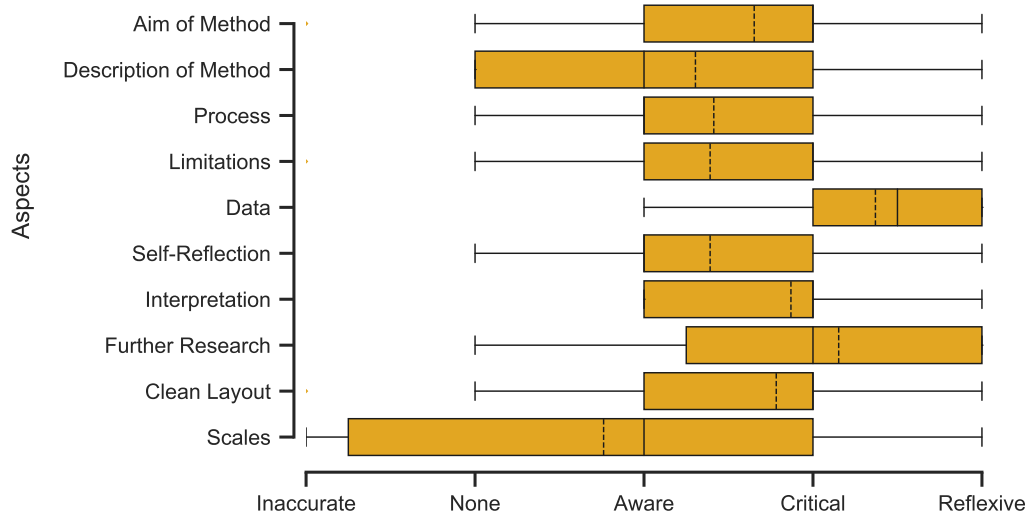


Fig. 4.2: Box plot of the results of the analysis using the analytical rubric. It shows the extent to which the authors of the articles point to an inaccurate, none, aware, critical or reflexive attitude on methodological aspects by applying principal component analysis and visualizing their data.

4.1.1 Data Reduction

The first challenge that will be discussed relates to the reduction of data, something that characterizes the use of principal component analysis, but which can have many consequences for the research in all respects – think of visualization, interpretation and drawing conclusions. As Kennedy et al. argue, researchers are convinced to "doing good with data"[3] when visualizing data, however, when certain choices or justifications are left there or misused, this can lead to inaccurate application of a certain method. One of the ideas behind data visualizations is that it makes data more accessible.[4] For PCA in particular, this involves

[3] Kennedy et al., “The work that visualisation conventions do,” 720.
[4] Hartmann, “Humanization of Knowledge through the Eye.”

a reduction in data. But what is the perspective or view that had been displayed and what consequences could have this for the original data? And can you say something about a data set in its whole, when only a part is displayed. A trustworthy approach to and accurate description of the data set is something that seems to be becoming increasingly important for the humanities, as this research shows.

Data What is striking about the results is the relatively high score for the aspect of *data*. Most of the data in the studies have been extensively and clearly explained and discussed, and ample time is devoted to the discussion of the data set. Almost all data sets consist of (spoken) texts, consisting of data in words, lemmas or sentences and derives from texts such as opera librettos, manuscripts, poetic texts, novels, plays, dialogues and dialects. The only non-lettered data consists for one article about newspaper layouts [ADHO_BKX] and for another article about pixels of heat images [DHQ_Sci]. In general, the researchers seem to have a reflexive and critical attitude towards the effect that PCA can have on their data. Limitations of the analysis that are mentioned relate in particular to the possible consequences for their data. Especially in authorship attribution studies one seems to be aware of the fact that PCA has a reducing effect and that the analysis is therefore more indicative than resulting and conclusive. However, in many articles insufficient information is given about the pre-processing steps needed to analyze the data using PCA. As an example of awareness of the reducing effect of PCA, some of the articles mention:

> [...] **some information is lost** when visualizing a high-dimensional data set in two or three dimensions. it is not 'a test of authorship but only of comparative resemblance [...] it is not 'a test of authorship but only of comparative resemblance'. Principal components is an unsupervised learning tool, **useful in discovering general clustering patterns** of the data. [LLC_SFRS, 75-76]

> It should be noted that 1500 word-samples run the **risk of increased imprecision**, a consideration which should nuance any interpretation of the results. [ADHO_G, 2]

Although the authors pay extensive attention to the discussion of the data set, and seem to be aware of the reducing effect, they do not discuss the possible manipulation, loss or absence of data for their research. In some cases, the number of variables in the data set has been reduced from 98, 326 or even 500 to just 2 variables and the researchers interpret these results as a whole of the data set and then take along their interpretations and conclusions further with them for the remainder of the study, whereas it is actually a reduced representation of the whole.

Aim of Method The considerations and argumentation of the choice to use PCA for the intended aim within the research has been examined, and also, to what extent the researchers seemed to express themselves critically or reflexively towards the application of this specific method. The findings based on the rubric show that a majority of the investigators only mentioned or briefly described the purpose and a definition of the method used, or didn't even mention them. A critical justification, explanation or discussion could only be found in a few articles. Further review of the articles shows that the main and most cited arguments for applying PCA were: reducing the dimensionality and (visually) identifying patterns,

clusters, similarities, differences or outliers. While, as mentioned above, it seems to be a challenge for Digital Humanities to apply a certain digital method in a well-considered way, for the application of PCA it seems generally known that it is best applied for indicative and exploratory purposes. In almost all articles words were used for the purposes of the method such as: explore, experiment, exploratory analysis, examine, pre-processing step, preliminary. This can be recognized in the following quotes:

> Principal Component models are used as **exploratory** tools to spot patterns in the data matrix. [DHQ_Sci, 7]
>
> PCA is often used to **examine** variation in language. [DSH_O, 641]
>
> [...] the kernel PCA sets up kernels and parameters more flexibly; thus, it is more suitable for use in **preliminary** and **exploratory** research [DHQ_SH, 2]
>
> Do these 24 documents really constitute one set?—were they translated by the same translators? To test this premise we employ a technique called principal component analysis (PCA) to **examine** the variation between documents. [LLC_HBW, 120]

Something that is mentioned by Rieder and Röhle as a risk for humanities research, is the application of PCA in order to reveal new perspectives, relationships or characteristics that have not been predicted or thought through beforehand.[5] In the articles that have been analyzed this does not appear to be a risk, but a valuable application of exploratory research. However, further critical reflection often seems to be lacking. The intrinsically value of exploration is also an important point of criticism on the application of digital methods in the humanities mentioned by Da. It is questionable whether in the humanities complex data can be reduced. According to her, patterns in complex data sets would not provide the answer to humanities questions, as she calls it: '[...] there is no rationale for such reductionism'.[6] However, when the research is exploratory in nature, PCA does not in a sense provide answers to concrete questions, but rather offers an opportunity to gain insight as a step towards concretion.

Explained Variance Another striking result is the explained variance of the PCA results of the studies, i.e. accounting the variability of the data. By means of PCA the number of variables in a study is reduced to principal components that explain almost all data and do not correlate with each other. The total variance of PCA indicates the percentage of variance explained by the components. In order to give practical meaning to the extracted factors, the aim is to explain the variance in the data as much as possible. In other disciplines there are certain guidelines for desirable values of explained variance. In natural sciences, the percentage or variance criterion implies that the total explained variance should be at least 95%, or that the last principal component explains a variance of more than 5%. In social sciences the guidelines are less sharp, as the data is often less exact. However, a total explained variance of at least 60% is desirable.[7] Looking at the corpus of this study, the average of the total variance explained by the components resulting from the PCA is 45,80%. Taking into account that we are dealing with other types of data, this is still much

5 Rieder and Röhle, "Digital methods: From challenges to Bildung."
6 Da, "The computational case against computational literary studies," 638.
7 Joseph F Hair et al., *Multivariate data analysis*, vol. 5, 3 (Prentice hall Upper Saddle River, NJ, 1998), 107.

lower than in other disciplines. The method of PCA, as mentioned earlier, is often used to investigate variations in, among other things, texts, words or novels. However, when the components explain a low percentage of the variance, this means that the components only explain to a limited extent the differences in the corpus. In some cases, the results of the PCA were considered valuable despite the low percentage of variance explained. The following examples show that although PCA hardly explains a thing, hypotheses are still being made:

> Although the percentage of variance explained by the PCA model is small (<10%), **no clustering is evident**, indicating the substantial homogeneity of librettos and theatrical plays. [DHS_ST, 415]

> An analysis based on the 600 most frequent words [...] displays **clear and distinct clustering** of the sections of all six texts, even though the first two principal components account for **only 42 percent of the total variance.** [LLC_H, 344]

In other cases, a more critical approach was taken towards the results with regard to the explained variance or the researchers explain why the results are valuable because of a high percentage:

> Projection onto the first, dominant, component in Fig. 10 hints that Hurlbert **may be 'closer'** to the Diary than Ward but we should note the **low percentage of variation explained (53.5%)** by looking at just two dimensions.[LLC_HC, 191]

> [...] the first two principal components **must be interpreted with caution**. [LLC_HH, 489]

> [...] we can see the cumulative proportion of variance of first two principal components surpasses 80%, and that of the first three components achieves 86%. So, the relationship of texts, obtained by PCA and seen from Figure 1, is **reliable**. [DHS_HJ, 360]

Even more important is that for 62 out of 141 visualizations, the variance explained by the components has not even been mentioned or discussed. The fact that PCA is a data-reducing analysis means that variance plays an important role in the interpretation of the results. The lack of variance, which is a key element of PCA, to some extent impairs the reliability of the studies. Clusters and outliers are detected, properties are assigned to variables and the visualizations are presented as evidence, without insight into the meaning of the components.

4.1.2 Transparency

Reporting the explained variance in research is also part of transparency. The use of digital methods goes hand in hand with the challenge of ensuring transparency within humanities research. Perils involve the complexity and invisibility of the processes of digital methods, tools, underlying technology and algorithms. Unawareness of the underlying operation of the methods and tools could result in a black-box.[8] The inscrutability and opacity of digital tools and methods require researchers to adopt an increasingly reflective and critical attitude towards their use. Therefore, the application of the PCA with respect to the process of implementation, the limitations of the method and to what extent the researchers showed some kind of self-reflection in the article were examined.

[8] Dobson, *Critical Digital Humanities: The Search for a Methodology*.

Process and Limitations The results show that in 10 out of 46 articles the process for implementing PCA is not or hardly discussed and that in 15 out of 46 articles some steps that have been taken were mentioned, but a critical or reflexive attitude seems often to be lacking. After a thorough description of the corpus and a brief explanation of PCA, the resulting visualizations are discussed abruptly in almost all cases. What is striking about this is the lack of a discussion of methodological choices, considerations and decisions made or taken during the analysis. In only a few articles limitations were pointed out with regard to the application of digital methods, e.g.:

> Nevertheless, exactly because **machines do not carry the same set of preconceptions as humans**, the application of stylometry is able to induce serendipity in humanities research and open up new perspectives. [ADHO_KSBW, 4]

> Within a data mining practice the information is processed in order to find **correlations rather than causations**. [DSH_Sci, 7]

In other scholarly fields, PCA is a method that has been used for much longer now and where a step-by-step explanation of the process may be less necessary. In the humanities, however, the use of this method is still relatively new, which makes it more important to describe and reflect on limitations and technological underpinnings of the method. As it is new it would make also sense to discuss and reflect its contribution to the field. Unless the researchers are aware of the poor results and want to disguise it. By making it clear that the researchers understand the methodological procedures, what the method does, how it works, what limitations but also what advantages the method has, the transparency of the research will be increased.

Tools Table 4.1 shows which tools, application software or programming languages have been used in the studies. Each tool requires certain settings, adjustments or preferences in the interface when used for computational analysis and visualization. Since the articles pay little – and in most cases even no – attention to this, it is questionable to what extent this is taken into account during their research. The lack of information on the use of which tool, which settings and which limitations the analysis contains, increases the black-box of technologies and digital methods in the humanities.[9] Although many Digital Humanities critics advocate an open source philosophy, to counteract the problem of black-boxing, the question is to what extent this has been institutionalized to this point. The results of this research show, therefore, that little attention is still being paid to this issue. An epistemological problem within Digital Humanities is to what extent researchers need to understand the technologies, tools and computational methods to use them.[10] Do you need to understand how a car works in order to use it? However, most of the critics in the debate argue that it is important to have precisely this knowledge about the methods they use and the underlying algorithms. One needs to know the possibilities, but also the limitations of a certain tool.[11] This can be broadened for the application of PCA, where both the method and the tool play an important role in this. As the results show, different tools are used for the

9 Dobson, *Critical Digital Humanities: The Search for a Methodology*.
10 Rieder and Röhle, "Digital methods: From challenges to Bildung."
11 Fred Gibbs and Trevor Owens, "Building better digital humanities tools," *DH Quarterly* 6, no. 2 (2012).

same method. One can understand the interface of a certain tool used, but without knowing exactly the aim and processes of the method, and vice versa. Another striking aspect of the results is that in 30 out of 46 articles the authors not even mentioned which tool or software was used for the principal component analyses. Only one article mentioned and discussed the algorithm behind PCA. Within the debate on the value of the use of digital tools and methods in the humanities, there is a lot of criticism about the lack of transparency and the risk of black-boxing. What we can at least learn from the results is that this transparency is indeed lacking in many articles. It is questionable to what extent researchers need to share their code, settings, algorithms. However, in order to ensure credibility, reliability and reproducibility, there is certainly a case for institutionalizing an open source philosophy.

Tab. 4.1: Overview of the tools, software application or programming languages used for PCA

Tool	Articles
Python	4
R	8
Excel Applet GenAIX	1
SPSS	1
Minitab Software	1
Evince by Predicte	1
Not mentioned	30

4.1.3 Objectivity

The use of digital methods aims to some extent to make research more objective than traditional interpretative research. In order to counteract the lure of objectivity, subjective choices therefore become all the more important. As mentioned earlier, applying quantitative research using computational tools goes hand in hand with making subjective choices and considerations. Numerous decisions are made about the selection of data, how the analysis is carried out and what is done with the results. It might becomes even easier to make methodological mistakes, because of the impressiveness of the amount of data and the many layers the research contains.[12] Therefore, this study examined the conditions in which PCA was applied and the extent to which decision making during application was substantiated in the study.

Scaled axes The most remarkable aspect of the results is the distribution of the appropriate use of the scales in the visualizations. The inaccurate use of scales on the axes is an example of manipulation with data that often occurs in PCA visualizations. The axes are often created automatically by the computational tools – in which certain default results and graphical displays are not adapted to the data and analysis – but the scaling of the axes is particularly important for the application of PCA.[13] The impact of inaccurate use of scaled axes is illustrated by skewed scaled axes relative to equally scaled axes in figure 4.3, using a simple but equal data set. It shows that with skewed axes, different clusters can be identified,

12 Rieder and Röhle, "Digital methods: From challenges to Bildung."
13 Chen et al., *Handbook of Data Visualization*.

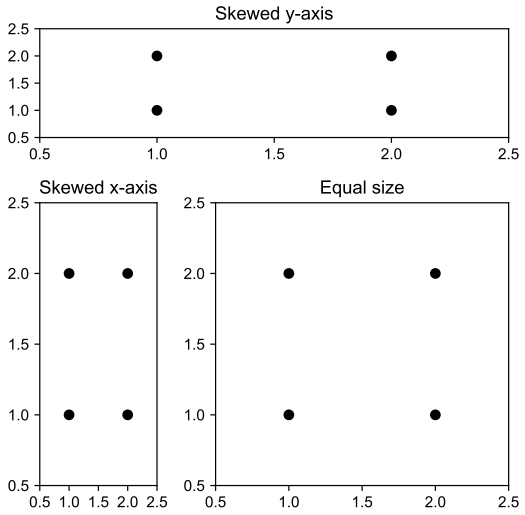
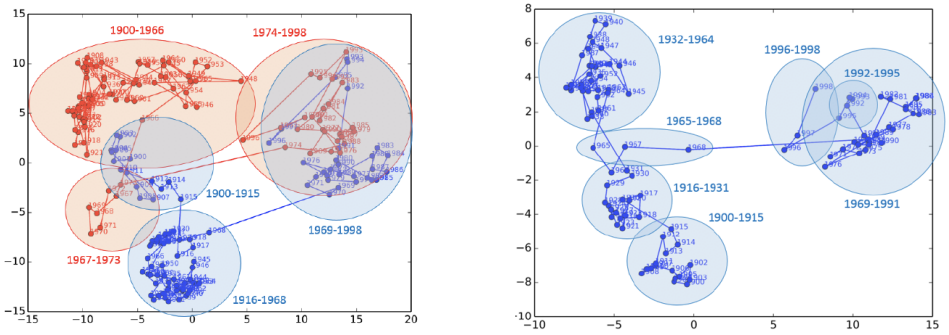


Fig. 4.3: Comparison of skewed and equal axes.

while the same data visualized with equal axes will not result in clusters. The use of skewed axes therefore leads to different results, which may lead to different interpretations and conclusions.[14] Examining the accuracy of the scales in the 46 analyzed articles, resulted in only 16 articles with equal scaled axes and 12 articles with skewed scales axes. In addition, 13 articles displayed the scales almost accurately, with the step size being the same on both axes but still distorted or different scaled axes in different visualizations within the same article (see figure 4.4). Finally, in 5 articles the PCA results were visualized in a scatter plot without scales (see figure 4.5). Inaccurate use of scaled axes may be a consequence of naive or incompetent use of the tools with which PCA has been applied and the results of which have been visualized. To investigate this further, I have tested the application of PCA using the most commonly used tools in the analyzed articles (in these cases programming languages R and Python, see table 4.1) to investigate the standard implied outcomes of these tools with default settings. As an example I used data from 50 irises from each of three varieties: iris setosa, iris versicolor, and iris virginica.[15] When PCA is applied in both tools, without any preference or adaptation to the scaling or axes settings, it illustrates the axes will turn out skewed by default. The results and code used for this can be seen in appendix 6.5. This research suggests that researchers often use the standard settings of a tool, while it can have major consequences for the interpretation of their results. Clusters, patterns and outliers can be incorrectly identified – which are basically the most important aspects of PCA – when the axes are not equally dispalyed.

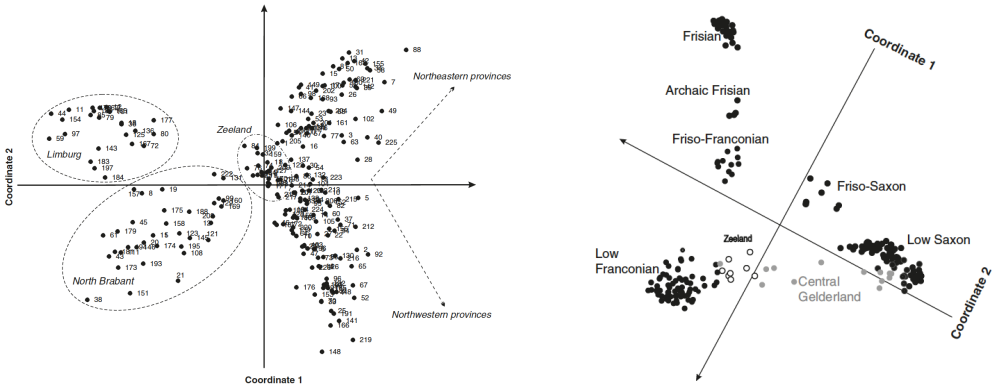
14 William S Cleveland, *Visualizing data* (Hobart Press, 1993).
15 Ibid.



(a) Accurate uses of scale

(b) Inaccurate uses of scale

Fig. 4.4: Examples of the use of different scaled axes in the same article. Figures adopted from the article: Buntinx, V., Kaplan, F., & Xanthos, A. (2017). *Layout analysis on newspaper archives.* [ADHO_BKX]



(a) Absence of scaling on the axes

(b) Rotated axes and absence of scaling on the axes

Fig. 4.5: Examples the use of axes without scaling. Figures adopted from the article: Manni, F., Heeringa, W., & Nerbonne, J. (2006). *To what extent are surnames words? Comparing geographic patterns of surname and dialect variation in the Netherlands.* [LLC_MHN]

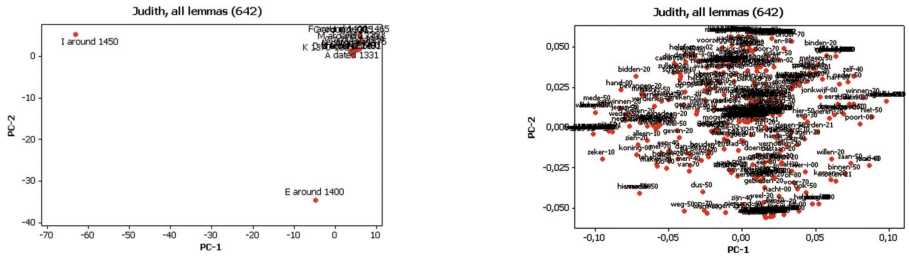


Fig. 4.6: Examples of unclear and messy labelling. Figures adopted from the article: van Dalen-Oskam, K. (2012). *The secret life of scribes. Exploring fifteen manuscripts of Jacob van Maerlant's Scolastica (1271)* [LLC_D]

Clean Layouts Data visualizations are formed on the basis of explicit and implicit conventions, which are needed to increase and strengthen the quality, objectivity and transparency of the figures and data.[16] Existing criticism of visualizations encompasses the idea that they favor certain points of view, strengthen existing and create new power relationships. On the other hand, creators of visualizations believe in the stimulating power of visualizations for a better understanding of data by making data accessible and transparent. Data visualizations may give the impression of showing the facts and reality through the intended objectivity and quality. However, a visualization is always a representation of reality.[17] Kennedy et al. mentioned in their article the importance of clean layouts, which would enhance the objectivity of the visualizations.[18] Tufte also argued the importance of uniformity in graphs to ensure good communication and perception of the information and he states: 'Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity'.[19] He mentioned the principle of using the highest possible data-ink ratio, which means that the 'ink' in a graphical representation of scientific information should only show the important data and associated information (legends, axes, etc.), without distracting and unnecessary elements.[20]

In this study the graphical display and layout of the resulting scatter plots obtained from the PCA has been examined, as it enhances the objectivity, clarity and legibility of the visualizations. Although a scatter plot is a fairly simple and straightforward graph, many of the visualizations analyzed made extensive use of different graphical shapes, colors and lines. The data points and interrelations are represented by geometric shapes and lines, which further simplifies the data.

16 Cf. Kennedy et al., "The work that visualisation conventions do"; Drucker, *Graphesis: Visual forms of knowledge production*; Tufte, *The visual display of quantitative information*; Tufte et al., *Visual explanations: Images and quantities, evidence and narrative*; Friendly and Denis, "The early origins and development of the scatterplot."
17 Kennedy et al., "The work that visualisation conventions do."
18 Ibid.
19 Tufte, *The visual display of quantitative information*, 56.
20 Tufte, "The visual display of quantitative information Graphics Press."

In some visualizations, the data points were difficult to see and the whole looked messy and sloppy, as can be seen in figure 4.6. In only a few cases a critical reflection was given, with regard to the graphical representation, e.g.:

> Each **circle in the plot represents** one play, and their relative proximity or distance indicates topic-based, thematic similarity or difference in the three dimensions shown. The **colors of the circles correspond** to the **conventional genre labels** of each play, which however do not influence the positions of the circles. The coloring only allows us to see to what degree the topic-based similarity of the plays corresponds with their conventional genre label.[DHQ_Sch, 14]

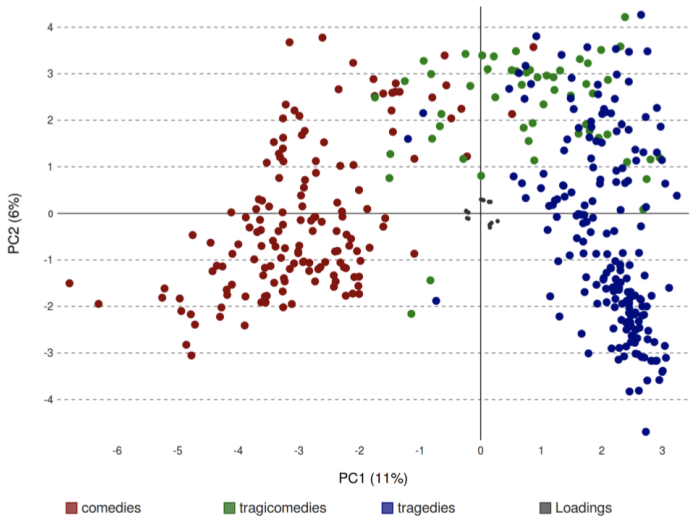


Fig. 4.7: Figure adopted from the article: Schöch, C. (2017). *Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama.*[DHQ_SCH]

In this article the graphical elements, such as the colors and the dots are argued and discussed. It provides a clear and readable picture and ensures that choices have been made carefully. Less clean layouts often contain unclear labels and messy ordering. In addition, they often lack a legend, color reasoning and axes. Other examples of clean and non-clean layouts can be seen in appendix 6.6.

4.1.4 Interpretation

The final challenge of using digital and computational methods that is central to this research concerns the interpretation of the results of PCA. The importance of interpretation in data visualizations is because " [...] visualizations are always interpretations — data does not have an inherent visual form that merely gives rise to a graphic expression."[21] And this is where visualization has come to play an important role. By arranging and displaying complex data in a certain way, it offers the possibility of analysis and interpretation to discover patterns and deviations.[22] When the process of visualization is sometimes seen as a method, it is presented another time as a result of a research. My analysis show that the authors of the

21 Drucker, *Graphesis: Visual forms of knowledge production*, 6.
22 Otty et al., "Data Visualization in the Humanities."

analyzed articles tend towards a critical and reflexive approach in their interpretation of the results and visualizations of PCA. Whereas the implementation of PCA and the process of analysis often seems to be discussed quickly and concisely, the results seem to come to most attention in extensive interpretations of the results and their corresponding visualizations. As discussed earlier, the interpretation of visualization is very important for the production of knowledge. In particular, the resulting scatter plots from PCA only show the results of the statistical analysis, without any context. In this way they are mainly interpreted, which is in line with the point of view of exploratory research. The interpretation of the visualizations is supported in the articles with words and expressions, such as: the results indicates, we observe, demonstrates discernible incoherence, suggests, is visibly different, show, represents, seem to provide some meaningful insights, clusters are noticeable. In the following quotations, the suggestive and observational way of interpretation and the importance of further interpretation of the visualizations becomes more clear:

> **Visual inspection reveals** three main clusters for each journal. Each of these clusters turns out to correspond to groups of clusters that has been detected in the previous projections. We **observe** that the layout of both journals has evolved in a similar way but with different timescales. [ADHO_BKX, 3]

> Both present visualisations in a way that **helps to think about the data** and both lead to **new ideas** as to what the next steps in the research could be. [LLC_D, 356]

> This diagram **shows an imperfect yet encouraging degree of separation**. The first, horizontal, component, clearly expresses a strong authorial signal. Taken together with the second, vertical, component, we find a nearly perfect separation of Paine, in the upper right-hand area of the graph, from the other authors. [DHQ_FH, 5]

> **More experiments are needed**, of course, to test the validity of these techniques beyond the domain of literature. [ADHO_R, 4]

The risk of visualizations is that it is assumed to be 'true' or 'factual', which tends more towards the original truth-to-nature principle of objectivity, whereas it always represents a representation of a certain perspective, depending on data, choices in analysis and manner of visualization. The idea of objectivity in the production of knowledge as we know it nowadays is always connected to human subjectivity.[23] Nevertheless, some visualizations and PCA results were analyzed as a form of evidence, using the following words, among others: is evidence, the reverse is true, it is clear that, without exception, one can distinguish discrete clusters. When visualizations are interpreted as objective evidence, it has a completely different meaning for the production of knowledge than when it is regarded as suggestive or observable. Drucker appropriately formulates this underlying risk of data visualizations: '[...] they are images that act as if they are just showing us what is, but in actuality, they are arguments made in graphical form'.[24] Because PCA is mainly used as preparatory and exploratory for follow-up analyses, it is important that the results are also interpreted in this way. And the value of PCA as a starting point for follow-up research is something the authors in the articles were very conscious, critical and reflective about.

[23] Daston and Galison, *Objectivity*.
[24] Drucker, *Graphesis: Visual forms of knowledge production*, 9-10.

Conclusion and Discussion

5

Over the last decade digital technology is changing the way in which people pursue research in the field of humanities. These developments of digital methods and innovative visualization techniques and their application in the humanities, have also opened up a critical debate on what these technologies contribute to the existing traditional methods, their promises, limitations and the way forward. This debate does not only affect the way research is conducted, but also the various effects of the epistemology and ontology of doing research within the Digital Humanities. In virtually all parts of research the digital transformation has consequences, e.g., transitions in the variety of subjects, changes in processes and analyses of the material, and effects on the interpretations of results. This thesis was inspired by a provoking article written by Da on the use of digital methods in the Digital Humanities.[1] According to her analysis, computational and quantitative research in the humanities lacks, among other things, statistical accuracy and justification of the method used. The debate poses the question whether the new methodological paradigm does not conflict with the traditional research in the humanities using close reading and hermeneutic interpretation.[2] In response to this debate, I have structurally analyzed the methodological aspects of processing and visualizing data using the principal component analysis (PCA) in 46 Digital Humanities studies, with the aim to answer the main question of this thesis: '*In what way is the principal component analysis methodologically applied and visualized within the scholarly field of Digital Humanities?*'. I have evaluated and analyzed the method section, the implementation and processing of PCA and the resulting visualizations of these studies.

More specifically, with this research I have concentrated on four major challenges of the use of digital methods that are very interconnected and noticeable within the Digital Humanities: 1) manipulation and reduction of data, 2) technological transparency of computational tools, 3) seduction of scientific objectivity and 4) the importance of interpretation of results and visualizations. The results provide insight into the extent to which the authors of the articles point to an awareness, a critical attitude or reflection on choices and actions of the aforementioned challenges of applying digital methods, with regard to the visualization of data, within their research. The main findings of these evaluations concern the lack of transparency in the implementation of the method and the perils of objectivity. The results show that a clear explanation, description or purpose of the use of PCA is often lacking and therefore tends towards naivety and ignorance regarding the critical epistemological discussion about the use of digital methods in humanities. PCA can be applied quite easily without paying excessive attention to the significance of the application for the data, results

1 Da, "The computational case against computational literary studies."
2 Owens, "Defining data for humanists: Text, artifact, information or evidence."

and interpretation. Most researchers seem to have an incomplete, reticent and closed position with respect to the methods. Especially for important elements of data analysis and visualization, such as explained variance, scaling of the axes and visual elements of graphical display, relevant information is frequently missing or the method is applied incorrectly.

The enthusiasm in which humanities, thwarted by hesitation, skepticism and criticism, have embraced the advent of digital methods seems to go hand in hand with inexperience and naivety. I don't want to claim that Digital Humanities scholars do not have the required knowledge and experience about digital methods, but, due there is the great lack of transparency. This lack of transparency also affects the reproducibility challenge Piper is referring to.[3] When certain choices, considerations and other aspects of the method would be discussed more explicitly and reflectively, research will become more reproducible. Due the very interpretative nature of the humanities, reproducibility is an issue that affects the discipline probably more when methods, i.e. digital analysis, that have limited transparency, are applied. If the procedures remain a black-box, they will stay invisible, complex and difficult to trace. Transparency contributes to the reliability and validity of the research and in order to obtain a more critical and better understanding of the results. Institutionalization of the above-mentioned open source philosophy of Digital Humanities scholars to counter the problem of black boxing is something that should receive more attention.

The interpretation of the visualizations seems to have effects on the lure of objectivity of academic research. Visual reasoning is more likely to be accepted as objective as it seems more credible than textual reasoning. It is the power of visualization that makes information visible and understandable through the use of visual communication. However, data visualizations are based on human subjectivity. Problems could be solved when scientific images are approached by logical processes and presented as a preliminary result of an ongoing process, open to interpretation. This would prevent presenting PCA visualizations as too definitive, with all the risks of over-interpretation. This marks also the importance of research into how technology is currently changing our understanding of what the humanities are. Drucker mentions, for example, that "qualitative judgment take priority over quantitative statements and presentations of facts'?"[4] It is questionable to what extent interpretation and qualitative judgment have changed the core principles of the humanities with digital, computational and quantitative methods. Although this may rather be something for follow-up research, this study makes clear that quantitative methods are increasingly taking part in research in the humanities, be it as exploratory preliminary research or as complementary to qualitative, more traditional, research. We should in any case continue to ask ourselves what Digital Humanities is about and what it means for the production of knowledge.

There has already been a plea for visual literacy for the Digital Humanities.[5] Concluding from the research in this thesis, the field would also benefit from digital literacy. This research concerned articles that mainly referred to statistical guidelines for the implementation of PCA, from the natural or social sciences. However, as discussed in the results, other data,

3 Piper, "Do we know what we are doing?"
4 Drucker, *Graphesis: Visual forms of knowledge production*, 6-7.
5 Jessop, "Digital visualization as a scholarly activity."

e.g. text-based objects and cultural artifices, are used. This confirms the need for guidelines, conventions and agreements for the application of digital methods, such as PCA, specifically suited to Digital Humanities research. Not only guidelines for digital methods, but also the embedded assumptions and concepts in computational tools used that are not always in accordance with the objectives of the humanities necessitate this. Programming languages, such as R and Python, which have mainly been used in this research, are complex programs for complicated codes that are not specifically designed for the humanities. It is important that digital methods are not simply adopted, but adapted to the humanities data and research that is applicable. Focusing on and reflecting on the application of R and Python and other software applications in Digital Humanities research, requires a tool-critical mindset of humanities scholars. This also applies to the visualization of data. Scientific graphical displays are mainly formed by the disciplines from which they originate, which may be at odds with the constructive principles of the humanities approaches to scholarly research. The question is whether these conventions and agreements can be used in interpretive research in the humanities. This critical reflection contributes to the key idea of the university, *Digital Bildung*.[6] One must continue to think critically and reflectively about the impact and process of digital practices and culture in a culture in a computational era avoiding technological enthusiasm.

As this research was exploratory in nature, it offers several possibilities for follow-up research. During the work, I tried to identify a representative sample of Digital Humanities research papers. Limited to this study was the domain definition of Digital Humanities in the scientific literature. A well-considered choice was made to select journals representing Digital Humanities, at the time of this project. As discussed in section 3, many Digital Humanities papers are also published in fields such as Computer Science or Computational Linguistic journals. In these other scholarly fields, other agreements and guidelines for publication may apply. It may therefore not be a perfect sample of the field, something that follow-up research could extend to a more diverse corpus. In addition, only the abstracts of the *Proceedings of the Annual Conference* of the ADHO have been analyzed. A limitation of this is that they are less complete in terms of methodological discussion than the full articles. Finally, I would like to flag the issue of subjectivity of the researcher. Although I conducted the data processing, analyses, etc. in a protocol based, systematic way, my own subjective interpretation of all the steps were part of this work, and therefore certainly has room for improvement. I think the annotation model and the analytical rubric have been a good starting point for more epistemological and critical research into the application of digital methods within the Digital Humanities.

In this research, clear things have emerged that are visible as right or wrong – think of the representation of the explained variance and the scaling of the axes – but due to the interpretative nature of the humanities, fixed guidelines are not necessarily feasible. A reliable, transparent, but above all reflective approach to the use of digital methods are becoming are becoming increasingly important, given the challenges facing Digital Humanities scholars, as this research shows.

6 Berry, "The computational turn: Thinking about the digital humanities."

Bibliography

Arnold, Taylor, and Lauren Tilton. “Distant viewing: analyzing large visual corpora.” *Digital Scholarship in the Humanities* 34, no. Supplement_1 (2019): i3–i16.

Beacham, Richard, Hugh Denard, Francesco Niccolucci, et al. “An introduction to the London charter.” *Joint Event CIPA/VAST/EG/EuroMed*, 2006.

Berry, David. “The computational turn: Thinking about the digital humanities.” *Culture machine* 12 (2011).

Berry, David M. “Introduction: Understanding the digital humanities.” In *Understanding digital humanities*, 1–20. Springer, 2012.

Borgman, Christine L. “The digital future is now: A call to action for the humanities,” 2010.

Burrows, John. “‘Delta’: a measure of stylistic difference and a guide to likely authorship.” *Literary and linguistic computing* 17, no. 3 (2002): 267–287.

Champion, Erik Malcolm. “Digital humanities is text heavy, visualization light, and simulation poor.” *Digital Scholarship in the Humanities* 32, no. suppl_1 (2016): i25–i32.

Chen, Chun-houh, Wolfgang Karl Härdle, and Antony Unwin. *Handbook of Data Visualization*. Springer, 2008.

Chen, Min, David Ebert, Hans Hagen, Robert S Laramee, Robert Van Liere, Kwan-Liu Ma, William Ribarsky, Gerik Scheuermann, and Deborah Silver. “Data, information, and knowledge in visualization.” *IEEE Computer Graphics and Applications* 29, no. 1 (2008): 12–19.

Cleveland, William S. *Visualizing data*. Hobart Press, 1993.

Da, Nan Z. “The computational case against computational literary studies.” *Critical Inquiry* 45, no. 3 (2019): 601–639.

Daston, Lorraine J, and Peter Galison. *Objectivity*. Zone Books, 2007.

Dobson, James E. *Critical Digital Humanities: The Search for a Methodology*. University of Illinois Press, 2019.

Drucker, Johanna. *Graphesis: Visual forms of knowledge production*. Harvard University Press, 2014.

———. “Humanities approaches to graphical display.” *Digital Humanities Quarterly* 5, no. 1 (2011): 1–21.

Eyers, Tom. “The Perils of the" Digital Humanities": New Positivisms and the Fate of Literary Theory.” *Postmodern Culture* 23, no. 2 (2013).

Few, Stephen, and Perceptual Edge. “Data visualization: past, present, and future.” *IBM Cognos Innovation Center*, 2007.

Friendly, Michael. “A brief history of data visualization.” In *Handbook of data visualization*, 15–56. Springer, 2008.

Friendly, Michael, and Daniel Denis. “The early origins and development of the scatterplot.” *Journal of the History of the Behavioral Sciences* 41, no. 2 (2005): 103–130.

Garcia Penalvo, Francisco J. “Digital Humanities Data Processing,” 2016.

Gibbs, Fred. “Digital humanities definitions by type.” In *Defining Digital Humanities*, 305–314. Routledge, 2016.

Gibbs, Fred, and Trevor Owens. “Building better digital humanities tools.” *DH Quarterly* 6, no. 2 (2012).

Gold, Matthew K. *Debates in the digital humanities*. U of Minnesota Press, 2012.

Hair, Joseph F, William C Black, Barry J Babin, Rolph E Anderson, Ronald L Tatham, et al. *Multivariate data analysis*. Vol. 5. 3. Prentice hall Upper Saddle River, NJ, 1998.

Hartmann, Frank. “Humanization of Knowledge through the Eye.” *Making Things Public–Atmospheres of Democracy*, 2005, 698–707.

Hayles, N Katherine. “How we think: Transforming power and digital technologies.” In *Understanding digital humanities*, 42–66. Springer, 2012.

Hill, Rosemary Lucy, Helen Kennedy, and Ysabel Gerrard. “Visualizing junk: Big data visualizations and the need for feminist data studies.” *Journal of Communication Inquiry* 40, no. 4 (2016): 331–350.

Holmes, David I. “The evolution of stylometry in humanities scholarship.” *Literary and linguistic computing* 13, no. 3 (1998): 111–117.

Jänicke, Stefan, Greta Franzini, Muhammad Faisal Cheema, and Gerik Scheuermann. “On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges.” In *EuroVis (STARs)*, 83–103. 2015.

———. “Visual text analysis in digital humanities.” In *Computer Graphics Forum*, 36:226–250. 6. Wiley Online Library, 2017.

Jessop, Martyn. “Digital visualization as a scholarly activity.” *Literary and Linguistic Computing* 23, no. 3 (2008): 281–293.

Jolliffe, Ian. *Principal component analysis*. Springer, 2011.

Jones, Steven E. *The Emergence of the Digital Humanities (Open Access)*. Routledge, 2013.

Kennedy, Helen, Rosemary Lucy Hill, Giorgia Aiello, and William Allen. “The work that visualisation conventions do.” *Information, Communication & Society* 19, no. 6 (2016): 715–735.

Khan, Muzammil, and Sarwar Shah Khan. “Data and information visualization methods, and interactive mechanisms: A survey.” *International Journal of Computer Applications* 34, no. 1 (2011): 1–14.

Kirschenbaum, Matthew. “What is digital humanities and what’s it doing in English departments.” *Debates in the digital humanities* 3 (2012).

Kirschenbaum, Matthew G. “What is digital humanities and what’s it doing in English departments?” In *Defining Digital Humanities*, 211–220. Routledge, 2016.

Liu, Alan. “The state of the digital humanities: A report and a critique.” *Arts and Humanities in Higher Education* 11, nos. 1-2 (2012): 8–41.

Moretti, Franco. *Distant reading*. Verso Books, 2013.

———. *Graphs, maps, trees: abstract models for a literary history*. Verso, 2005.

Münster, Sander, and Melissa Terras. “The visual side of digital humanities: a survey on topics, researchers, and epistemic cultures.” *Digital Scholarship in the Humanities*, 2019.

Otty, Lisa, Tara Thomson, and Tara Thomson. “Data Visualization in the Humanities.” *Research Methods for Creating and Curating Data in the Digital Humanities*, 2016.

Owens, Trevor. “Defining data for humanists: Text, artifact, information or evidence.” *Journal of Digital Humanities* 1, no. 1 (2011): 6–8.

Pandey, Anshul Vikram, Anjali Manivannan, Oded Nov, Margaret Satterthwaite, and Enrico Bertini. “The persuasive power of data visualization.” *IEEE transactions on visualization and computer graphics* 20, no. 12 (2014): 2211–2220.

Piper, Andrew. “Do we know what we are doing?” *Journal of Cultural Analytics*, 2019.

———. “There Will be Numbers Journal of Cultural Analytics.” 2016. Accessed September 12, 2019. `culturalanalytics.org/2016/05/there-will-be-numbers/`.

Porsdam, Helle. “Too much ‘digital’, too little ‘humanities’? An attempt to explain why many humanities scholars are reluctant converts to digital humanities,” 2011.

Rieder, Bernhard, and Theo Röhle. “Digital methods: From challenges to Bildung,” 2017.

Schöch, Christof. “Big? smart? clean? messy? Data in the humanities,” 2013.

Schreibman, Susan, Ray Siemens, and John Unsworth. “The digital humanities and humanities computing: An introduction.” *A companion to digital humanities*, 2004, xxiii–xxvii.

Sinclair, Stéfan, Stan Ruecker, and Milena Radzikowska. “Information visualization for humanities scholars.” *Literary Studies in the Digital Age-An Evolving Anthology*, 2013.

Stone, Maureen. “Challenge for the Humanities.” *Working together or apart: Promoting the next generation of digital scholarship*, 2009, 43.

Sula, Chris Alen. “Digital humanities and libraries: A conceptual model.” *Journal of Library Administration* 53, no. 1 (2013): 10–26.

Svensson, Patrik. “The landscape of digital humanities.” *Digital Humanities*, 2010.

Tufte, Edward R. *The visual display of quantitative information*. Vol. 2. Graphics press Cheshire, CT, 2001.

———. “The visual display of quantitative information Graphics Press.” *Cheshire, Connecticut*, 1983.

Tufte, Edward R, Susan R McKay, Wolfgang Christian, and James R Matey. *Visual explanations: Images and quantities, evidence and narrative*, 1998.

Underwood, Ted. *Distant horizons: digital evidence and literary change*. University of Chicago Press, 2019.

Vanhoutte, Edward. *Defining digital humanities: a reader*. Ashgate Publishing, Ltd., 2013.

———. “The gates of hell: History and definition of digital| humanities| computing.” In *Defining Digital Humanities*, 135–172. Routledge, 2016.

Wold, Svante, Kim Esbensen, and Paul Geladi. “Principal component analysis.” *Chemometrics and intelligent laboratory systems* 2, nos. 1-3 (1987): 37–52.

List of Figures

List of Tables

List of Listings

Appendices

6

6.1 Appendix 1: Corpus

Tab. 6.1: Overview of the selected articles

Authors	Title	Year	Journal	Label
Akihiro, K.	Regional Classification of Traditional Japanese Folk Songs from Southwest Regions	2017	Proceedings of the Digital Humanities Congress	ADHO_A
Bonch-Osmolovskaya, A., & Skorinkin, D.	Text mining War and Peace: Automatic extraction of character traits from literary pieces.	2016	Digital Scholarship in the Humanities	DSH_BS
Bruster, D., & Smith, G.	A new chronology for Shakespeare's plays.	2014	Digital Scholarship in the Humanities.	DSH_BS
Buntinx, V., & Kaplan, F.	Negentropic linguistic evolution: A comparison of seven languages	2018	Proceedings of the Digital Humanities Congress	ADHO_BK
Buntinx, V., Kaplan, F., & Xanthos, A.	Layout analysis on newspaper archives	2017	Proceedings of the Digital Humanities Congress	ADHO_BKX
Burrows, J.	Who wrote Shamela? Verifying the authorship of a parodic text.	2005	Literary and linguistic computing	LLC_B
Connors, L.	Function word analysis and questions of interpretation in early modern tragedy	2008	Proceedings of the Digital Humanities Congress	ADHO_C
De Gussem, J.	A Stylometric Study of Nicholas of Montiéramey's Authorship in Bernard of Clairvaux's Sermones de Diversis	2017	Proceedings of the Digital Humanities Congress	ADHO_G
Dixon, P., Mannion, D., & Burgess, W. G.	Johnson,'Misargyrus', and Richard Bathurst.	2018	Digital Scholarship in the Humanities	DSH_DM
Eder, M.	Does size matter? Authorship attribution, small samples, big problem.	2013	Digital Scholarship in the Humanitie	DSH_E
Forstall, C. W., Jacobson, S. L., & Scheirer, W. J.	Evidence of intertextuality: investigating Paul the Deacon's Angustae Vitae.	2011	Literary and linguistic computing	LLC_FJS
Forsyth, R., & Holmes, D	The Writeprints of Man: a Stylometric Study of Lafayette's Hand in Paine's 'Rights of Man'	2018	Digital Humanities Quarterly	DHQ_FH
Gladwin, A. A., Lavin, M. J., & Look, D. M.	Stylometry and collaborative authorship: Eddy, Lovecraft, and 'The Loved Dead'.	2017	Digital Scholarship in the Humanities	DSH_GLL
Hawkins, L. F.	Computational Models for Analyzing Data Collected from Reconstructed Cuneiform Syllabaries	2018	Digital Humanities Quarterly	DHQ_H
Holmes, D. I., & Crofts, D. W.	The diary of a public man: a case study in traditional and non-traditional authorship attribution.	2010	Literary and linguistic computing	LLC_HC
Hoover, D. L.	Multivariate analysis and the study of style variation.	2003	Literary and linguistic computing	LLC_H
Hoover, D. L., & Hess, S.	An exercise in non-ideal authorship attribution: The mysterious Maria Ward.	2009	Literary and linguistic computing	LLC_HH
Hou, R., & Jiang, M.	Analysis on Chinese quantitative stylistic features based on text mining.	2014	Digital Scholarship in the Humanities	LLC_HJ
Hulle, D. & Kestemost, M.	Stylochronometry and the Periodization of Samuel Beckett's Prose	2016	Proceedings of the Digital Humanities Congress ADHO_HK	

Continued on next page

Tab. 6.1 – *Continued from previous page*

Authors	Title	Year	Journal	Label
Hung, J. J., Bingenheimer, M., & Wiles, S	Quantitative evidence for a hypothesis regarding the attribution of early Buddhist translations.	2009	Literary and linguistic computing	LLC_HBW
Hyvönen, S., Leino, A., & Salmenkivi, M.	Multivariate analysis of Finnish dialect data—an overview of lexical variation.	2007	Literary and linguistic computing	LLC_HLS
Iyeiri, Y., Yaguchi, M., & Baba, Y.	Principal component analysis of turn-initial words in spoken interactions.	2011	Literary and linguistic computing	LLC_IYB
Ji, M.	A corpus-based study of lexical periodization in historical Chinese.	2010	Literary and linguistic computing	LLC_J
Jockers, M. L., & Witten, D. M.	A comparative study of machine learning methods for authorship attribution.	2010	Literary and linguistic computing	LLC_JW
Kestemont, M., Moens, S., & Deploige, J.	Collaborative authorship in the twelfth century: A stylometric study of Hildegard of Bingen and Guibert of Gembloux.	2013	Digital Scholarship in the Humanities	DSH_KMD
Kestemont, M., Stronks, E., de Bruin, M., & de Winkel, T.	Did a Poet with Donkey Ears Write the Oldest Anthem in the World? Ideological Implications of the Computational Attribution of the Dutch National Anthem to Petrus Dathenus	2017	Proceedings of the Digital Humanities Congress	ADHO_KSBW
Lucic, A., & Blake, C.	Comparing the Similarities and Differences between Two Translations	2011	Proceedings of the Digital Humanities Congress	ADHO_LB
Manni, F., Heeringa, W., & Nerbonne, J.	To what extent are surnames words? Comparing geographic patterns of surname and dialect variation in the Netherlands.	2006	Literary and linguistic computing	LLC_MHN
Manousakis, N., & Stamatatos, E.	Devising Rhesus: A strange 'collaboration'between Aeschylus and Euripides.	2017	Digital Scholarship in the Humanities	DSH_MS
Merriam, T.	Untangling the derivatives: points for clarification in the findings of the Shakespeare Clinic.	2009	Literary and linguistic computing	LLC_M
Oakes, M. P.	Computer stylometry of CS Lewis's The Dark Tower and related texts.	2017	Digital Scholarship in the Humanities	DSH_O
Rajan, V.	Quantifying scripts: Defining metrics of characters for quantitative and descriptive analysis.	2016	Digital Scholarship in the Humanities	DSH_R
Reeve, J. P.	Does "Late Style" Exist? New Stylometric Approaches to Variation in Single-Author Corpora	2018	Proceedings of the Digital Humanities Congress	ADHO_R
Reynolds, N. B., Schaalje, G. B., & Hilton, J. L.	Who wrote Bacon? Assessing the respective roles of Francis Bacon and his secretaries in the production of his English works	2012	Literary and linguistic computing	LLC_RSH
Rosen-Zvi, M., et al.	Stranger Genres: Computationally Classifying Reprinted Nineteenth Century Newspaper Texts	2018	Proceedings of the Digital Humanities Congress	ADHO_RZ
Saccenti, E., & Tenori, L.	Multivariate modeling of the collaboration between Luigi Illica and Giuseppe Giacosa for the librettos of three operas by Giacomo Puccini.	2014	Digital Scholarship in the Humanities	DSH_ST
Schaalje, G. B., Fields, P. J., Roper, M., & Snow, G. L.	Extended nearest shrunken centroid classification: a new method for open-set authorship attribution of texts of varying sizes.	2011	Literary and linguistic computing	LLC_SFRS
Schöch, C.	Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama	2017	Digital Humanities Quarterly	DHQ_Sch
Sciuto, C.	Recording invisible proofs to compose stone narratives. Applications of Near Infrared Spectroscopy in provenance studies.	2018	Digital Humanities Quarterly	DHQ_Sci

Continued on next page

Tab. 6.1 – *Continued from previous page*

Authors	Title	Year	Journal	Label
Skorinkin, D.	Character-distinguishing features in fictional dialogue: the case of War and Peace	2017	Proceedings of the Digital Humanities Congress	ADHO_S
Suzuki, T., & Hosoya, M.	Computational Stylistic Analysis of Popular Songs of Japanese Female Singer-songwriters	2014	Digital Humanities Quarterly	DHQ_SH
Takahashi, M., Tezuka, K. & Yano, T.	Identifying the author of the Noh play by considering a rhythmic structure — Validating the application of multivariate analysis	2013	Proceedings of the Digital Humanities Congress	ADHO_TTY
Temple, J.T.	A Multivariate Synthesis of Published Platonic Stylometric Data	1996	Literary and Linguistic computing	LLC_T
Uesaka, A., & Murakami, M.	Authorship problem of Japanese early modern literatures in Seventeenth Century	2013	Proceedings of the Digital Humanities Congress	ADHO_UM
Uesaka, A., & Murakami, M.	Verifying the authorship of Saikaku Ihara's work in early modern Japanese literature; a quantitative approach	2014	Digital Scholarship in the Humanities	DSH_UM
van Dalen-Oskam, K.	The secret life of scribes. Exploring fifteen manuscripts of Jacob van Maerlant's Scolastica (1271).	2012	Literary and Linguistic Computing	LLC_D
Wieling, M., Shackleton Jr, R. G., & Nerbonne, J.	Analyzing phonetic variation in the traditional English dialects: Simultaneously clustering dialects and phonetic features.	2013	Literary and Linguistic Computing	LLC_N

6.2 Appendix 2: Analytical Rubric

Tab. 6.2: Analytical Rubric: Operationalisation and implementation of PCA

Aspects	Reflexive (3)	Critical (2)	Aware (1)	None (0)	Inaccurate (-1)
Aims and Methods	Aim of the method has been critically explained Method is clearly explained, argued and criticized	Aim of the method has been explained Method is clearly explained	Aim of the method is only defined Definition of the method is mentioned	Aim of the method is not appointed Definition of the method is not appointed	Inaccurate aim of the method Method is inaccurate explained
Process	Process is clearly explained, argued and criticized	Process is clearly explained	Process is mentioned	Process is not appointed	Process is inaccurate explained
Limitations	Limitations are clearly explained, argued and criticized	Critical arguments on limitations of the method	Some limitations are mentioned	Limitations of the method are not appointed	Inaccurate reflection on the limitations
Data	Dataset is clearly explained and argued	Dataset is explained	Dataset is defined	Dataset is mentioned	Dataset contains mistakes and is used inaccurate
Self-reflection	Researchers are very self-reflexive of the choices and actions that have been made	Researchers are critical of the choices and actions that have been made	Researchers are aware of the importance of a critical attitude	Researchers are not self-reflective or critical	Researchers are inaccurate reflexive
Interpretation	Results are critical interpreted and argued	Results are a bit critical interpreted	Results are interpreted, but lack in reflexivity	Results are presented as objective facts	Results are inaccurate interpreted
Further Research	Further research has been critically argued and applied	Further research has been critically argued	Further research has been mentioned	Further research has not been mentioned	Further research has been suggested inaccurately
Graphical Display	Visualizations are very clear and concise, supplemented with legend or other information	Visualizations have a clear layout and are easy to read	Visualizations are readable, but lack a clear graphic layout	Visualizations are cluttered and hard to read	Visualizations are inaccurate displayed
Scales	Scales are accurate used and displayed	Scales are close to accurate used and displayed	Scales are wrong used, but displayed	There are no scales displayed	Scales are inaccurate displayed

6.3 Appendix 3: Annotation models

Tab. 6.3: Annotation Model Articles

Category	Aspect	Label	Question
General	Title	title	What is the title of the article?
	Authors	authors	Who are the authors of the article?
	Year	year	From which year is the article?
	University	university	From which university are the authors?
	Subject	university	What is the subject of the article?
	Study	university	What does the study of the article contain?
	Visualizations	visualizations_total	How many visualizations does the article contain?
		visualizations_pca	How many PCA visualizations does the article contain?
Aims and Methods	Transparency of method	pca_definition	How is PCA defined in the article?
		pca_advantages	Which advantages of PCA are mentioned?
		pca_limitations	Which limitations of PCA are mentioned?
		pca_ref	Which reference if used to explain PCA?
	Argumentation	goal	What is the explained goal of using PCA?
	Combination	combination	Is the PCA in combination with other analyses?
	Tools	tools	Which tool is used to perform the analysis?
	Accessibility	accessibility	Is the data accessible for others?

Tab. 6.4: Annotation Model Visualizations

Category	Aspect	Label	Question
Principal Component Analysis	Datatype	datatype	What kind of data does the dataset contain?
	Variables	variables	How many variables does the dataset contain?
	Dimensions	dimensions	To how many dimensions is the data reduced?
	Variance	variance_[total, pc1, pc2, ...]	How many of the variance does the data account for?
Graphical Display	Title	title	What is the title of the visualization?
	Caption	caption	What is the caption of the visualization?
	Section	section	In which section are the visualizations?
	Scales	scales	Are the scales well used?
	Color	color	Which colors are used?
		color_ref	What does the use of color represent?
	Labels	labels	Does the visualization contain labels and to what do the labels refer?
	Shapes	shapes	Does the visualization contain shapes and to what do the shapes refer?
	Clusters	clusters	Are there any visuals used to define the clusters?
	Legend	legend	Does the visualization contains a legend and to what do the legend refer?

6.4 Appendix 4: Results of the Analytical Rubric

Tab. 6.5: Results of the Analytical Rubric

Article ID	Aim of Method	Description	Process	Limitations	Data	Self-Reflection	Interpretation	Further Research	Clean Layout	Scales	Total
ADHO_A	1	0	1	0	2	0	2	3	2	-1	10
ADHO_BK	1	0	0	1	1	1	1	1	2	1	9
ADHO_BKX	2	1	1	1	2	2	3	3	3	2	20
ADHO_C	2	3	2	1	3	2	1	3	3	1	21
ADHO_G	2	2	1	2	3	2	2	1	3	1	19
ADHO_HK	1	0	1	2	3	2	1	3	1	0	14
ADHO_KSBW	1	0	0	2	3	2	2	2	2	1	15
ADHO_LB	1	0	0	0	1	0	1	1	2	-1	5
ADHO_R	1	0	0	0	2	0	1	1	1	1	7
ADHO_RZ	2	0	0	1	1	0	2	1	2	3	12
ADHO_S	1	0	1	1	1	1	2	3	2	1	13
ADHO_TTY	0	0	1	0	2	1	1	1	1	2	9
ADHO_UM	1	0	0	0	2	1	1	1	1	2	9
DHQ_FH	3	2	2	2	3	2	2	3	2	1	22
DHQ_H	3	3	2	3	3	2	3	2	3	2	26
DHQ_Sch	3	2	2	2	3	2	3	3	3	2	25
DHQ_Sci	2	1	2	2	3	3	1	1	2	-1	16
DHQ_SH	2	1	1	1	3	2	2	1	2	2	17
DHS_BOS	1	0	0	0	1	0	1	0	2	2	7
DHS_BS	2	1	1	1	3	1	2	2	3	3	19
DHS_E	3	1	1	2	3	2	2	2	1	1	18
DHS_HJ	2	2	1	1	1	1	1	2	1	-1	11
DHS_KMD	2	3	2	2	3	2	2	3	3	2	24
DHS_MS	2	3	3	2	3	1	2	3	1	2	22
DHS_R	0	1	0	1	2	0	1	2	1	-1	7
DHS_ST	2	3	2	2	3	1	3	2	3	-1	20
DHS_UM	2	1	2	1	2	1	2	3	2	2	18
DSH_DMB	2	2	3	2	3	2	3	3	2	2	24
DSH_GLL	3	1	2	2	3	3	2	3	2	0	21
DSH_O	3	2	1	1	3	1	2	2	2	1	18
LLC_B	1	0	1	1	2	1	3	3	2	2	16
LLC_D	2	2	3	0	3	1	2	3	0	-1	15
LLC_FJS	-1	0	0	-1	2	1	1	2	2	-1	5
LLC_H	-1	0	0	2	2	1	1	2	1	1	9
LLC_HBW	2	2	3	2	3	2	2	3	2	1	22
LLC_HC	2	2	3	3	2	2	2	3	2	-1	20
LLC_HH	2	1	3	2	2	2	3	2	2	2	21
LLC_HLS	2	3	3	2	2	2	2	2	2	0	20
LLC_IYB	1	1	2	2	3	2	3	1	1	-1	15
LLC_JW	2	2	1	2	2	2	2	3	2	1	19
LLC_Mea	1	2	1	2	2	1	2	3	1	-1	14
LLC_MHN	2	2	2	2	3	1	2	3	-1	0	16
LLC_RSH	1	2	2	1	3	1	2	1	2	1	16
LLC_SFRS	3	2	1	2	3	3	2	3	2	-1	20
LLC_T	2	2	3	2	2	1	2	2	2	2	20
LLC_WSN	2	2	2	2	2	1	1	2	0	0	14

6.5 Appendix 5: PCA with Different Tools

6.5.1 PCA with Python

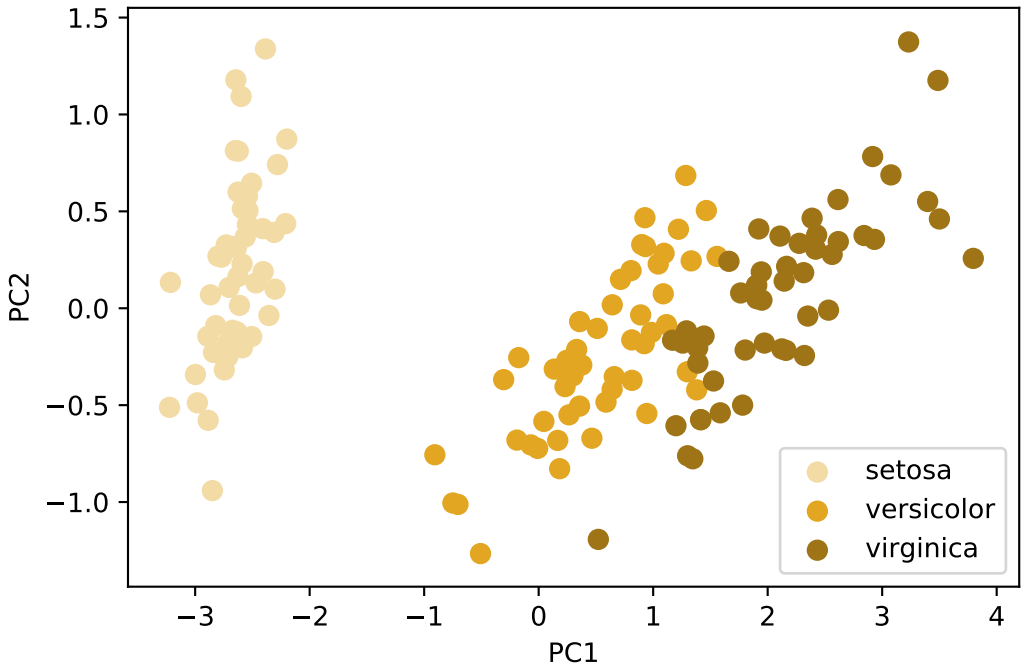


Fig. 6.1: PCA using Matplotlib in Python

```
import matplotlib.pyplot as plt
from sklearn import datasets
from sklearn.decomposition import PCA

df = datasets.load_iris()
x = df.data
y = df.target
target_names = df.target_names

pca = PCA(n_components=2)
x_r = pca.fit(x).transform(x)

plt.figure()
colors = ['#f3dba6','#e2a622', '#9E7417']

for color, i, target_name in zip(colors, [0, 1, 2], target_names):
plt.scatter(x_r[y == i, 0], x_r[y == i, 1], color=color, label=target_name)

plt.xlabel('PC1',fontsize=10)
plt.ylabel('PC2',fontsize=10)
plt.legend(loc='best', scatterpoints=1)
```

Listing 6.1: PCA code lines using Python

6.5.2 PCA with R

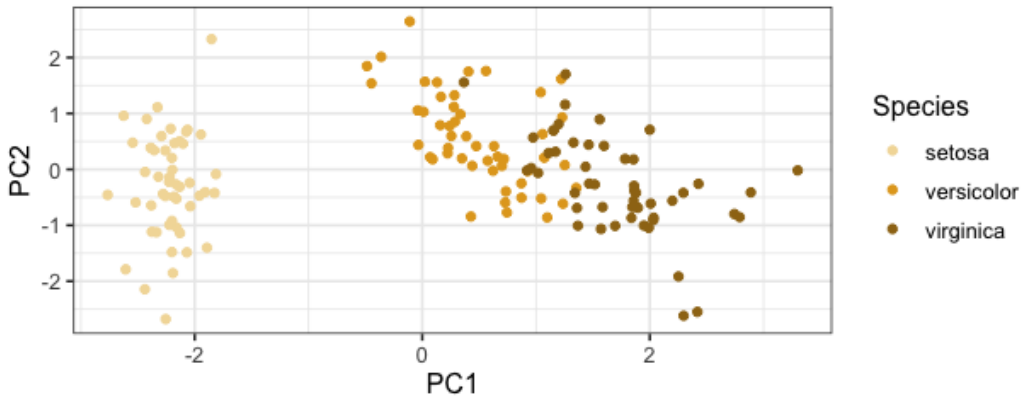


Fig. 6.2: PCA using ggplot in R

```
library(ggplot2)
library(cowplot) # multiple plots in a grid
library(dplyr)

head(iris)

iris %>% select(-Species) %>% # remove Species column
scale() %>%                   # scale to 0 mean and unit variance
prcomp() ->                   # do PCA
pca                           # store result as 'pca'

head(pca$x)

pca_data <- data.frame(pca$x, Species=iris$Species)
head(pca_data)

ggplot(pca_data, aes(x=PC1, y=PC2, color=Species)) + geom_point()
```

Listing 6.2: PCA code lines using R

6.6 Appendix 6: Graphical Displays of PCA Visualizations

6.6.1 Clean Layouts of PCA Visualizations

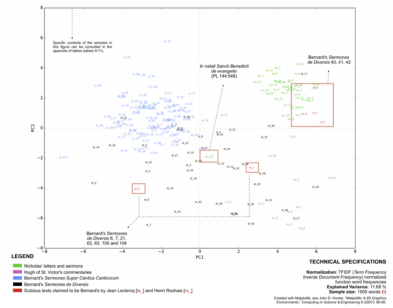


Fig. 6.3: De Gussem, J. (2017) *A Stylometric Study of Nicholas of Montiéramey's Authorship in Bernard of Clairvaux's Sermones de Diversis* [ADHO_G]

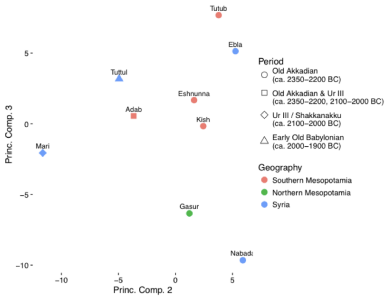


Fig. 6.4: Hawkins, L. F. (2018) *Computational Models for Analyzing Data Collected from Reconstructed Cuneiform Syllabaries.* [DHQ_H]

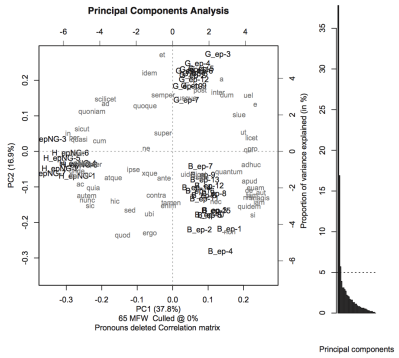


Fig. 6.5: Kestemont, M., Moens, S., & Deploige, J. (2013) *Collaborative authorship in the twelfth century* [DSH_KMD]

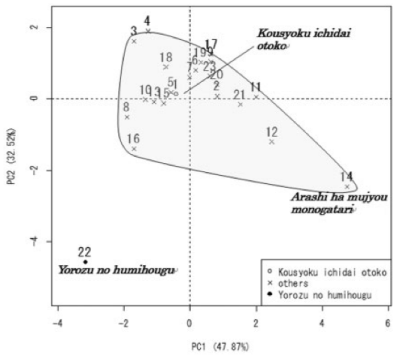


Fig. 6.6: Uesaka, A., & Murakami, M. (2014) Verifying the authorship of Saikaku Ihara's work in early modern Japanese literature [DSH_UM]

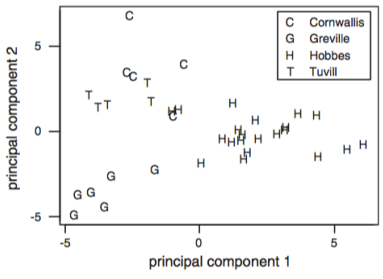


Fig. 6.7: Reynolds, N. B., Schaalje, G. B., & Hilton, J. L. (2012) *Who wrote Bacon? Assessing the respective roles of Francis Bacon and his secretaries in the production of his English works* [LLC_RSH]

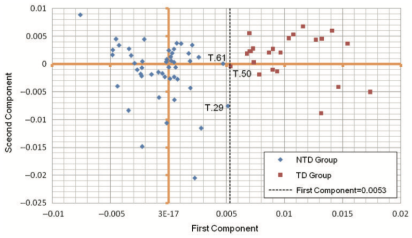


Fig. 6.8: Hung, J. J., Bingenheimer, M., & Wiles, S (2009) *Quantitative evidence for a hypothesis regarding the attribution of early Buddhist translations.* [LLC_HBW]

6.6.2 Less Clean Layouts of PCA Visualizations

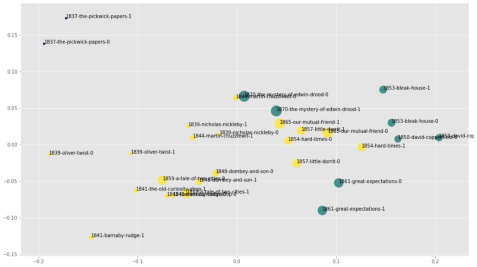


Fig. 6.9: Reeve, J. P. (2018) *Does "Late Style" Exist? New Stylometric Approaches to Variation in Single-Author Corpora* [ADHO_O]

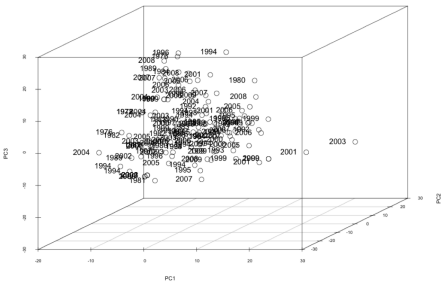


Fig. 6.10: Suzuki, T., & Hosoya, M. (2014) *Computational Stylistic Analysis of Popular Songs of Japanese Female Singer-songwriters* [DHQ_SH]

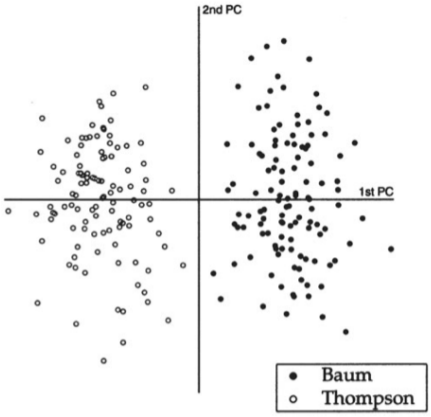


Fig. 6.11: Gladwin, A. A., Lavin, M. J., & Look, D. M. (2017) *Stylometry and collaborative authorship: Eddy, Lovecraft, and 'The Loved Dead'.* [DSH_GLL]

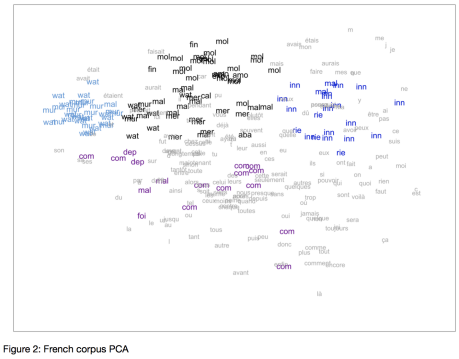


Fig. 6.12: Hulle, D. & Kestemost, M. (2016) *Stylochronometry and the Periodization of Samuel Beckett's Prose* [ADHO_HK]

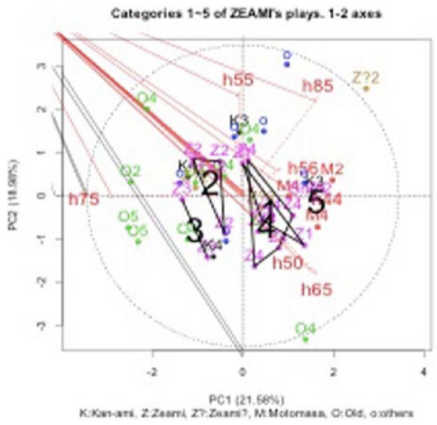


Fig. 6.13: Takahashi, M., Tezuka, K. & Yano, T. (2013) *Identifying the author of the Noh play by considering a rhythmic structure.* [ADHO_TTY]

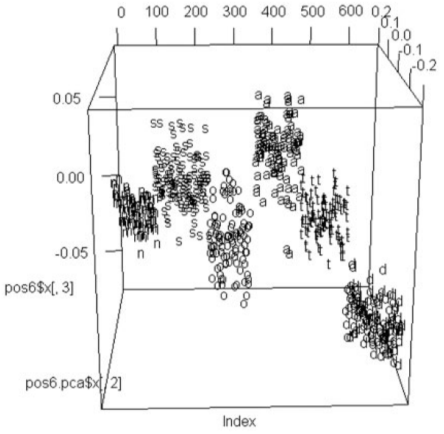


Fig. 6.14: Hou, R., & Jiang, M. (2014) *Analysis on Chinese quantitative stylistic features based on text mining.* [DSH_HJ]