

SILKE SCHWANDT (ED.)

---

# DIGITAL METHODS IN THE HUMANITIES

---

CHALLENGES, IDEAS, PERSPECTIVES

DIGITAL HUMANITIES RESEARCH  
**BIELEFELD** UNIVERSITY PRESS

Silke Schwandt (ed.)  
Digital Methods in the Humanities

## Editorial

Digital Humanities is an evolving, cross cutting field within the humanities employing computer based methods. Research in this field, therefore, is an interdisciplinary endeavor that often involves researchers from the humanities as well as from computer science. This collaboration influences the methods applied as well as the theories underlying and informing research within those different fields. These implications need to be addressed according to the traditions of different humanities' disciplines. Therefore, the edition addresses all humanities disciplines in which digital methods are employed. **Digital Humanities Research** furthers publications from all those disciplines addressing the methodological and theoretical implications of the application of digital research in the humanities. The series is edited by Silke Schwandt, Anne Baillot, Andreas Fickers, Tobias Hodel and Peter Stadler.

**Silke Schwandt** (Prof. Dr.), born 1980, teaches Digital and Medieval History at Bielefeld University. She received her PhD in Medieval History from Goethe-University Frankfurt am Main in 2010. Her research focus in Digital History lies with the transformation of scholarly practices through digitalization and with the advancement of digital literacy.

Silke Schwandt (ed.)

# **Digital Methods in the Humanities**

Challenges, Ideas, Perspectives

**[transcript]**



This volume has been prepared within the framework of the Collaborative Research Center SFB 1288 "Practices of Comparing. Ordering and Changing the World", Bielefeld University, Germany, funded by the German Research Foundation (DFG).



### **Bibliographic information published by the Deutsche Nationalbibliothek**

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <http://dnb.d-nb.de>



This work is licensed under the Creative Commons Attribution 4.0 (BY) license, which means that the text may be remixed, transformed and built upon and be copied and redistributed in any medium or format even commercially, provided credit is given to the author. For details go to <http://creativecommons.org/licenses/by/4.0/>

Creative Commons license terms for re-use do not apply to any content (such as graphs, figures, photos, excerpts, etc.) not original to the Open Access publication and further permission may be required from the rights holder. The obligation to research and clear permission lies solely with the party re-using the material.

© **Silke Schwandt (ed.)**

© **First published in 2021 by Bielefeld University Press, an Imprint of transcript Verlag**

<http://www.bielefeld-university-press.de>

Cover layout: Maria Arndt, Bielefeld  
Proofread by Julia Becker, Bielefeld  
Typeset by Michael Rauscher, Bielefeld  
Printed by Majuskel Medienproduktion GmbH, Wetzlar  
Print-ISBN 978-3-8376-5419-6  
PDF-ISBN 978-3-8394-5419-0  
<https://doi.org/10.14361/9783839454190>

Printed on permanent acid-free text paper.

# Contents

---

## Introduction

Digital Humanities in Practice <i>Silke Schwandt</i> .....	7
---	---

## I. Challenges for the Humanities

### Open Access, Open Data, Open Software?

Proprietary Tools and Their Restrictions <i>Helene Schlicht</i> .....	25
--	----

### Navigating Disciplinary Differences in (Digital) Research Projects Through Project Management

<i>Anna Maria Neubert</i> .....	59
---------------------------------	----

## II. From Text to Data

### From Text to Data

Digitization, Text Analysis and Corpus Linguistics <i>Patrick Jentsch, Stephan Porada</i> .....	89
--	----

## III. Digital Research Perspectives from Different Humanities Disciplines

### Testing Hypotheses with Dirty OCR and Web-Based Tools in Periodical Studies

<i>Malte Lorenzen</i> .....	131
-----------------------------	-----

### **Challenging the *Copia***

Ways to a Successful Big Data Analysis of Eighteenth-Century  
Magazines and Treatises on Art Connoisseurship

*Joris Corin Heyder* ..... 161

### **Text Mining, Travel Writing, and the Semantics of the Global**

An AntConc Analysis of Alexander von Humboldt's  
*Reise in die Aequinoktial-Gegenden des Neuen Kontinents*

*Christine Peters* ..... 185

### **From Serial Sources to Modeled Data**

Changing Perspectives on Eighteenth-Century Court Records  
from French Pondicherry

*Anna Dönecke* ..... 217

### **Looking for Textual Evidence**

Digital Humanities, Middling-Class Morality, and the Eighteenth-Century  
English Novel

*Ralf Schneider, Marcus Hartner, Anne Lappert* ..... 239

### **The Historical Semantics of Temporal Comparisons Through the Lens of Digital Humanities**

Promises and Pitfalls

*Michael Götzelmann, Kirill Postoutenko, Olga Sabelfeld, Willibald Steinmetz* ..... 269

**Authors** ..... 309

# Introduction

## Digital Humanities in Practice

---

Silke Schwandt<sup>1</sup>

### Digital methods for humanities research: chances and challenges

Digital Humanities (DH) is a growing field within the Humanities dealing with the application of digital methods to humanities research on the one hand as well as addressing questions about the influence of digital practices on research practices within the different humanities disciplines on the other. Edward Vanhoutte differentiates between computing methods being used “for and in the humanities”.<sup>2</sup> In his view the field of Digital Humanities, which was referred to as Humanities Computing before 2004, profited from the fact that the development of the first electronic computers were well underway during the Second World War, but were only fully operational after the war was over. This meant that their original military purpose, primarily in the field of ballistics and cryptanalysis, became obsolete, and the developers involved started looking for new operational scenarios in which the computers could be put to use. This failure, as Vanhoutte puts

---

1 I want to thank all contributors to this volume for their articles as well as their patience and dedication during our collaboration. The same goes for all members of team INF without whom this volume would not have been possible. This is especially true for Julia Becker, who proofread this volume and made it into what it is today. This book has been written within the framework of the Collaborative Research Center SFB 1288 “Practices of Comparing. Changing and Ordering the World”, Bielefeld University, Germany, funded by the German Research Foundation (DFG).

2 Vanhoutte, Edward, *The Gates of Hell: History and Definition of Digital | Humanities | Computing*, in: Melissa M. Terras/Julianne Nyhan/Edward Vanhoutte (eds.), *Defining Digital Humanities: A Reader*, London: Routledge, 2016, 119–156, 120.

it, allowed the computers to be used in the field of the humanities, especially in machine translation, from the 1950s onwards.<sup>3</sup> Clearly, this marks the beginning of the use of computing *for* the humanities, rather than *in* the humanities. Although Vanhoutte argues that both aspects can never be fully separated from each other, most digital practices can usually be attributed more to the one than to the other. At first glance, automatic word-by-word translations seem to be the attempt to use the computer in a clearly framed environment where the researchers trusted that its abilities would do exactly what they expected. It was only when the automatic translations started to provide unexpected results that the researchers started to think about their perception and understanding of the – in this case English – language while looking for explanations for the mistakes the computer made. Vanhoutte refers to Roberto Busa who “identified the major problem with research in Machine Translation not as the inadequacy of computers to deal with human language, but as man’s insufficient comprehension of human languages”.<sup>4</sup> Busa himself is one of the earliest and most important pioneers in Humanities Computing, or Digital Humanities, since he started a cooperation with IBM in order to create a concordance of the works of St. Thomas Aquinas in the 1940s. His relatively early assessment demonstrates the impact that the use of computational, or digital, methods can have on our understanding of the humanities as a research field and on the objects of that research. Busa hints at the necessity to alter our conceptions of language rather than looking for computational miscalculations. It is this impact that substantiates the apprehension that the field of Digital Humanities (or Humanities Computing as it was called during his time) is not only an advanced methodology but a research field in its own right. The vastness of such a field that might encompass any digital practices in the humanities – from communication practices to data management and data mining – accounts for the lack of a formal definition of what Digital Humanities actually is. The website “What is Digital Humanities” alone offers 817 different definitions collected by the

---

3 E. Vanhoutte, *The Gates of Hell*, 120–123.

4 E. Vanhoutte, *The Gates of Hell*, 125. Vanhoutte refers to *Busa, Roberto*, *The Annals of Humanities Computing: The Index Thomisticus*, in: *Computers and the Humanities* 14 (1980), 83–90, 86.

project “Day of DH” from 2009 to 2014.<sup>5</sup> Helene Schlicht and Anna Maria Neubert offer more insight into the definitions, workings, and self-determinations of Digital Humanities in their respective contributions to this volume.<sup>6</sup>

The historical account of Edward Vanhoutte shows one thing for sure that is also present in most Digital Humanities definitions: DH is a genuinely interdisciplinary endeavor. It brings together two very distinct research areas, Computer Science and the humanities, as well as many diverse research disciplines, methods and questions. The productive interaction between computer scientists and humanities researchers is one of the biggest chances and at the same time the biggest challenge in DH. As shown in the example from the early days of automated text analysis, the use of computational methods can inspire new research in the humanities. Unfortunately, their implementation is also often seen as an unnecessary and time consuming undertaking that only reproduces results that could have been generated by ‘traditional’ methods as well.<sup>7</sup> This impression leads to the assumption that DH is merely about methodology and focuses too much on the digital side of things, highlighting the results rendered by the application of digital tools to (mostly) text material. The innovative potential of interdisciplinary research of this kind is easily overlooked and downplayed. While it is absolutely necessary that research projects in DH offer interesting perspectives for both Computer Science and the humanities, the tendency to overemphasize the value of the new and advanced computer technologies belittles the importance of the humanities. Regardless of the alleged progress that comes with digitalization or the supposedly higher objectivity inherent in empirical data, it is still necessary and will remain essential to interpret the results produced by computational methods to arrive at reliable propositions.

---

5 Heppner, Jason, What Is Digital Humanities, <https://whatisdigitalhumanities.com/> [accessed: 21.08.2019].

6 See the contributions of Helene Schlicht and Anna Maria Neubert in this volume.

7 See for a similar discussion Schwandt, Silke, Digitale Objektivität in Der Geschichtswissenschaft? Oder: Kann Man Finden, Was Man Nicht Sucht?, in: Rechtsgeschichte – Legal History 24 (2016), 337–338. doi:10.12946/rg24/337-338.

## 1. Doing DH in Bielefeld: data infrastructure and Digital Humanities

In 2017, the German Research Foundation (DFG) granted funding to the Collaborative Research Center (SFB) “Practices of Comparing, Ordering and Changing the World”.<sup>8</sup> The Research Center consists of fourteen individual subprojects led by researchers from many different humanities disciplines, such as History, Literary Studies, Art History, Political Science, and Law. Situated at the heart of the center is the infrastructural project INF “Data Infrastructure and Digital Humanities” which is “responsible for supervising all data- and information-related activities by providing a collaborative digital work and research environment for the whole SFB.”<sup>9</sup> The project comprises expertise from the field of computer and information science as well as from the humanities, thus being well positioned to advise the other subprojects and to further the development of digital methods for the humanities. The main trajectories of the INF project include the implementation of a communication and project management tool for the Research Center as well as a data publication platform, where all historical source material is made available in digital formats. These aspects belong to the field of Research Data Management. Additionally, INF also supports the researchers in all questions regarding the use of digital methods for their subprojects. After developing a workflow for the digitization of documents with the help of OCR tools,<sup>10</sup> we advised six projects in total on how to tackle their research interests by using digital methods. They come from a variety of humanities disciplines and used different tools and analytical methods.

At the core of our work lies the task of modeling.<sup>11</sup> The practice of modeling may not be totally unknown to humanities scholars, although it has not yet been extensively discussed as such. Modeling seems to belong to the Sciences and has long been described as one of their core scholarly practices – especially in Physics. The need to implement the practice of modeling into

---

8 *Universität Bielefeld*, SFB 1288, Practices of Comparing, Ordering and Changing the World, [https://www.uni-bielefeld.de/\(en\)/sfb1288/](https://www.uni-bielefeld.de/(en)/sfb1288/) [accessed: 21.08.2019].

9 *Universität Bielefeld*, SFB 1288, TP INF, Data Infrastructure and Digital Humanities, [https://www.uni-bielefeld.de/\(en\)/sfb1288/projekte/inf.html](https://www.uni-bielefeld.de/(en)/sfb1288/projekte/inf.html) [accessed: 21.08.2019].

10 This workflow is described in detail in the contribution to this volume by Patrick Jentsch and Stephan Porada.

11 Anna Maria Neubert describes our work in detail in her contribution to this volume.

the humanities comes from the wish to productively interact with computational methods. Digital tools need a model to work with, an explicit and consistent representation of the world. Humanities researchers may have such representations at hand for the time periods, societies, etc., which they regard as their research objects. But they seldom frame them as a model. Willard McCarty defines such models as “either a *representation of something for purposes of study*, or a *design for realizing something new*.”<sup>12</sup> In our work at the Research Center we learned that modeling in order to build a representation for purposes of study is essentially a process of translation and transformation. It requires a great deal of communication and mutual understanding. Working in the humanities calls for adaptable interpretations that form, for example, our narrations of the past. Computer scientists, on the other hand, are trained to solve problems by finding one answer to any question. Therefore, the process of modeling does pose a challenge, especially to the humanities researcher. But it also opens up new ways of interacting with our knowledge about our research material and questions. McCarty points out two effects of computing to that end: “first, the computational demand for tractability, i. e. for complete explicitness and absolute consistency; second, the manipulability that a digital representation provides”.<sup>13</sup> In my opinion, it is the second effect, the manipulability of digital representations that offers the most interesting possibilities for the humanities. After using one distinct, explicit, and consistent model to arrive at that representation, the interpreter can always go back and change his or her presuppositions. Often, the digital representation that offers ways of manipulation is realized through visualizations.<sup>14</sup> These can be graphs, diagrams, trees, or network visualizations. Martyn Jessop sees the strength of digital tools of visualization in “[...] the ability of these tools to allow visual perception to be used in the creation or discovery of new knowledge.”<sup>15</sup> He stresses that in using visualization tools knowledge is not only “transferred, revealed, or perceived, but is created through a dynamic process.”<sup>16</sup> He also claims that “[digital visualiza-

---

12 McCarty, Willard, *Humanities Computing*, Houndmills: Palgrave Macmillan, 2014, p. 24.

13 W. McCarty, *Humanities Computing*, 25.

14 Jessop, Martyn, *Digital Visualization as a Scholarly Activity*, in: *Literary and Linguistic Computing* 23 (2008), 281–293. doi:10.1093/lc/fqn016.

15 M. Jessop, *Digital Visualization as a Scholarly Activity*, 282.

16 M. Jessop, *Digital Visualization as a Scholarly Activity*, 282.



tion] allows manipulation of both the graphical representation and the data it is derived from.”<sup>17</sup> Therefore, each visualization represents a certain interpretation of the source data, which depends on a manipulated version of that data. Bettina Heintz, a German sociologist working on the epistemological challenges posed by scientific visualizations, discusses the practice of such manipulations as one of the central practices in working with digital tools. The information behind the visualization is “altered, filtered, smoothed, and adjusted, until there is a relation between the expected and the presented”.<sup>18</sup> This practice does not only happen at the beginning of the research process but also over and over again during the research process. Interacting with digital tools in this way is a “genuinely experimental process”.<sup>19</sup> As McCarty says, “modelling problematizes”.<sup>20</sup> Hence, through visualization, the process of modeling can be continuously reevaluated. Modeling, as well as visualizing, enables humanities researchers to explore their digitalized source material in new ways. “As a tool of research, then, modelling succeeds intellectually when it results in failure, either directly within the model itself or indirectly through ideas it shows to be inadequate.”<sup>21</sup> What McCarty calls ‘failure’ could also be framed as ‘productive irritation’ – something that irritates the expectations of the researchers, which differs from their previous knowledge in such a way that it inspires new ideas about the allegedly well-known material.<sup>22</sup>

Six of the individual research projects in the Research Center at Bielefeld University have taken up this challenge and decided to evaluate digital methods for their humanities research. They joined the team of project INF in modeling their research ideas so that we could find digital tools that would help to answer those questions. In line with the overall research interests

---

17 M. Jessop, *Digital Visualization as a Scholarly Activity*, 238.

18 Heintz, Bettina/Huber, Jörg, *Der verführerische Blick: Formen und Folgen wissenschaftlicher Visualisierungsstrategien*, in: Bettina Heintz/Jörg Huber (eds.), *Mit dem Auge denken: Strategien der Sichtbarmachung in wissenschaftlichen und virtuellen Welten* (Theorie:Gestaltung 01), Zürich/Wien/New York: Voldemeer; Springer, 2001, 31.

19 B. Heintz/J. Huber, *Der verführerische Blick*, 23.

20 W. McCarty, *Humanities Computing*, 26.

21 W. McCarty, *Humanities Computing*, 26.

22 See Schwandt, Silke, *Digitale Methoden Für Die Historische Semantik: Auf Den Spuren Von Begriffen In Digitalen Korpora*, in: *Geschichte und Gesellschaft* 44 (2018), 107–134 for the idea of such productive irritation.

of the SFB 1288, these research questions all focus on practices of comparing while addressing such practices in different times, different genres or media, and performed by different historical actors. Practices of comparing seem to be ubiquitous – even today. What makes them historically interesting are the situational contexts in which they are being used, where they either stabilize certain ideas and structures or re-organize and change them. Comparing the modern West to the rest of the world, generating narratives of supremacy or eurocentrism, seems almost natural. The analysis of the emergence and the development of this specific comparison as well as the careful scrutiny of the situations in which this comparison is being made offer new insights into the development of nation states, of racism, and much more.<sup>23</sup> Digital tools of annotation and text analysis have proven to be especially useful in supporting research into practices of comparing since they allow, for example, simultaneous viewing of results as well as the detection of speech patterns representing specific modes of comparing. At the same time, DH methods are themselves often comparative and, therefore, implementing them makes it imperative to reflect on our own practices of comparing.<sup>24</sup>

## 2. Matching research practices and digital tools

The research projects, which serve as the basis for the contributions to this volume, all deal with textual material. It was therefore necessary to find tools for automatic textual analysis that would match the different underlying research questions. As text analysis tools we decided to work with *Voyant Tools* and *AntConc*. They both offer ample possibilities to calculate word frequencies, compile concordances, among other things, as well as provide visualizations of patterns within text documents or corpora.

---

23 Epple, Angelika/Erhart, Walter, Die Welt beobachten – Praktiken des Vergleichens, in: Angelika Epple/Walter Erhart (eds.), *Die Welt beobachten – Praktiken des Vergleichens*, Frankfurt/New York: Campus, 2015, 7–31.

24 Neubert, Anna/Schwandt, Silke, Comparing in the Digital Age. The Transformation of Practices, in: Angelika Epple/Walter Erhart /Johannes Grave (eds.): *Practices of Comparing. Towards a New Understanding of a Fundamental Human Practice*. Bielefeld 2020 [in print].

Voyant Tools is a web platform containing several open access text analysis tools.<sup>25</sup> It was developed by Geoffrey Rockwell and Stéfán Sinclair and is freely accessible on the web. The tools available operate mainly on word frequencies as well as the calculations of word distances. They span from well-known applications such as word cloud visualizations (*Cirrus*) to more elaborate tools focusing on the calculation of word repetitions throughout a text (*Knots*).<sup>26</sup> For the purposes of the projects in this volume the scope of tools provided by Rockwell and Sinclair is enough. In practice, it seems to be especially appealing to literary scholars and their interests in the use, frequency, and distribution of words and phrases throughout a text. Malte Lorenzen makes use of Voyant Tools in his article “Testing Hypotheses with Dirty OCR and Web-Based Tools in Periodical Studies”.<sup>27</sup> One of the tools he uses is *Cirrus*, the word cloud tool. Although the developers claim that word clouds “are limited in their interactivity [...] [and] do not allow exploration and experimentation”,<sup>28</sup> Lorenzen uses a series of these clouds to achieve just that. Confronting the different clouds with each other renders them exploratory after all through the practice of comparing. At the center of this comparison lies data that can be viewed as a representation of text, or rather as information about text. Rockwell and Sinclair claim that, in general, “[v]isualizations are transformations of text that tend to *reduce* the amount of information presented, but in service of drawing attention to some significant aspect.”<sup>29</sup> In the case of the word cloud the ‘significant aspect’ is the frequency of words in relation to each other represented by the relative size of their visualization. Hence, using digital text analysis tools often does not give us concrete or direct information about texts as a whole but about words, or character combinations, that need to be related to textual documents as superordinated, larger units before they can be interpreted. As Rockwell and Sinclair put it, “the magic of digital texts is that they are composed of discrete units of information – such as the character unit – that can be infinitely

---

25 Rockwell, Geoffrey/Sinclair, Stéfán, *Voyant*. See through your Text, <https://voyant-tools.org/> [accessed: 27.08.2019].

26 Rockwell, Geoffrey/Sinclair, Stéfán, *Tools*, <https://voyant-tools.org/docs/#!/guide/tools> [accessed: 27.08.2019].

27 See the contribution of Malte Lorenzen in this volume.

28 G. Rockwell/S. Sinclair, *Text Analysis and Visualization*, 276.

29 G. Rockwell/S. Sinclair, *Text Analysis and Visualization*, 276. Highlights in the original.

reorganized and rearranged on algorithmic whims”.<sup>30</sup> Whether it is magical or not, analyzing small, linguistic units of information instead of reading texts as indivisible entities offers new insights for researchers working on textual material as is being proven by the contributions in this volume. Joris C. Heyder and Christine Peters made use of a tool called AntConc for the same purpose.<sup>31</sup> Developed by Laurence Anthony,<sup>32</sup> AntConc “is a freeware, multiplatform tool for carrying out corpus linguistics research and data-driven learning”.<sup>33</sup> Other than Voyant Tools, it is a stand-alone tool that can be downloaded and installed locally on a computer. The tool comprises a Concordance Tool, a Concordance Plot Tool, which offers a barcode visualization of a keyword in context results, a File View Tool, N-Grams and Collocates Tools as well as Word List and Keyword List Tools. This range of tools is especially useful for studies interested in the word use present in certain documents or corpora. It offers the possibility to look for words surrounding specific keywords that offer insight into the concepts represented by words.

The contributors to this volume used digital text analysis tools such as Voyant Tools and AntConc in order to explore new ways to analyze the material they were researching. Rockwell and Sinclair describe two principles that they deem important when engaging with automatic text analysis: “Don’t expect too much from the tools [and] [t]ry things out”.<sup>34</sup> The first is about perspective. “Most tools at our disposal have weak or nonexistent semantic capabilities; they count, compare, track, and represent words, but they do not produce meaning – we do.”<sup>35</sup> While it seems obvious that the count of words does not carry semantic meaning, it is necessary to keep it in mind while looking for hooks for interpretation. This is also what makes working in DH a challenge. It is imperative to learn how to read visualizations and data as well as we read text. “Visualizations make use of a visual

---

30 G. Rockwell/S. Sinclair, *Text Analysis and Visualization*, 279.

31 See their contributions in this volume.

32 Anthony, Laurence, AntConc Homepage, <https://www.laurenceanthony.net/software/antconc/> [accessed: 27.08.2019].

33 Anthony, Laurence, AntConc (Windows, Macintosh OS X, and Linux), <https://www.laurenceanthony.net/software/antconc/releases/AntConc358/help.pdf> [accessed: 27.08.2019], 1.

34 G. Rockwell/S. Sinclair, *Text Analysis and Visualization*, 288.

35 *Ibid.*, 288.

grammar, just as language requires a linguistic grammar, and we need to be able to parse what we see before attempting to analyze and understand it [...].”<sup>36</sup> This is exactly why DH is a genuinely interdisciplinary endeavor making use of two things: digitization, or technologization, and hermeneutic interpretation. New digital technology transforms how we perceive and store information. It changes the ways of (social) interaction and communication. It allows access to vast amounts of information that need new ways of organization. And although these new technologies seem to be constantly evolving and becoming more and more important, it is equally important to make sense of these changes, to gain a new perspective, and to stay in touch with these developments in order to maintain a grip on them. In short: “[A]s digital technologies become increasingly pervasive, the work and skills of Digital Humanists become increasingly important.”<sup>37</sup>

### 3. Digital research perspectives in the humanities

While it seems to be almost impossible to separate computing *for* the humanities from computing *in* the humanities, the contributions in this volume focus on the implementation of digital methods in different humanities disciplines. By discussing the chances and challenges posed by this methodological endeavor, the contributors also touch on questions of the impact that working with digital tools has on the research practices of their respective fields. Their contributions are accompanied by three articles written by members of the project team INF trying to frame the setting of our collaborative work at Bielefeld University.

Helene Schlicht and Anna Maria Neubert deal with two important aspects of the general setup of our collaborative work within the Research Center in their respective articles. Helene Schlicht focuses on questions of “Open Source, Open Data, and Open Software”. She analyzes the “role of Open Science in the research landscape of the humanities in general and DH in particular”.<sup>38</sup> At present, questions of open access play a prominent role in

---

<sup>36</sup> *Ibid.*, 287.

<sup>37</sup> *M. Terras*, *Peering inside the Big Tent*, 270.

<sup>38</sup> See Schlicht, “Open Access, Open Data, Open Software? Proprietary Tools and Their Restrictions” in this volume.

political discussions about and within the humanities. In DH the implementation of open science solutions is much farther along. Schlicht argues that the contention of the two fields might help the advancement of both them. One of the problems she points out is the possible conflict between disciplines in DH. Anna Maria Neubert also discusses chances and challenges of interdisciplinary work in her contribution explicitly focusing on “Navigating Disciplinary Differences [...] Through Project Management”.<sup>39</sup> While it is not specific to DH projects, project management certainly helps with their organization and execution. It is especially important to take into account the possibly different research interests of the disciplinary groups participating in the projects as well as the different pace in research and publication. Neubert also discusses most of the software tools we used for the organizational side of our collaboration.

In their contribution on “From Text to Data. Digitization, Text Analysis, and Corpus Linguistics”,<sup>40</sup> Patrick Jentsch and Stephan Porada describe the technical workflows that we implemented for the collaboration. The main piece of the article deals with the digitization pipeline that was used to render the historic source material machine readable. Including this article into the volume demonstrates how important it is to include computer scientists into DH teams and also to take their research interests seriously. Only then does the collaboration rise to its full potential. It is also elementary to a volume focusing on digital methods to be transparent about every part of those methods and give credit where credit is due.

The contributions in this volume come from the fields of Computer Science, History, Literary Studies, and Art History. They represent the different approaches to research, different views and takes on text and interpretation.

One of the biggest challenges for the implementation of digital methods is the availability of digital source material – especially for historically oriented projects. Malte Lorenzen’s contribution deals with the chances offered and challenges posed by dirty OCR as a means to test the efficiency of digital methods for periodical studies from a Literary Studies’ point of

---

39 See Neubert, “Navigating Disciplinary Differences in (Digital) Research Projects Through Project Management” in this volume.

40 See Jentsch and Porada, “From Text to Data. Digitization, Text Analysis, and Corpus Linguistics” in this volume.

view. In his own words, his article has “experimental character”<sup>41</sup> and shows how exploring digital tools can further humanities research. He argues for a combination of close and distant reading that is necessary to integrate both quantitative digital methods and hermeneutic methods in the humanities, which is a position that can be found in many of the articles. Similar in the general trajectory of his interest in the chances and challenges posed by methods of Optical Character Recognition (OCR) and its use for historically oriented research is Joris C. Heyder’s article on “Challenging the *Copia*”.<sup>42</sup> He, also, wants to analyze great amounts of data, which is why a well-functioning OCR is crucial. While Malte Lorenzen uses the digital toolkit Voyant Tools to look for single terms and their usage in his material, Joris C. Heyder uses AntConc and its Concordance Tool to sort through the available material in search for the most interesting texts, building a corpus for his analysis from there.<sup>43</sup> Both articles use what we would call big data, but with different research questions and assumptions. Both explore the data with the help of digital tools arriving at different conclusions since they address the data on different levels – Lorenzen looks at the lexical level, whereas Heyder concentrates on the document level. Comparing the two articles demonstrates the manifold applications of digital methods in the humanities. What they have in common is the interest in “quick and dirty” digitization as a means to sort through large amounts of data.<sup>44</sup> They go about this task by testing the hypotheses they already have in mind after using traditional hermeneutic methods in designing their projects. Christine Peters follows a similar approach in her article on Alexander von Humboldt and his travel writings.<sup>45</sup> Alexander von Humboldt is probably one of the most well-known historical figures in world literature, and beyond. Christine Peters takes on the task of trying to find new perspectives on his travel writings in her contribution.

---

41 See Lorenzen, “Testing Hypotheses with Dirty OCR and Web-Based Tools in Periodical Studies” in this volume.

42 See Heyder, “Challenging the *Copia*. Ways to a Successful Big Data Analysis of Eighteenth-Century Magazines and Treatises on Art Connoisseurship” in this volume.

43 See the discussion of these tools above.

44 See Heyder, “Challenging the *Copia*. Ways to a Successful Big Data Analysis of Eighteenth-Century Magazines and Treatises on Art Connoisseurship” in this volume.

45 See Peters, “Text Mining, Travel Writing, and the Semantics of the Global. An AntConc Analysis of Alexander von Humboldt’s *Reise in die Aequinoktal-Gegenden des Neuen Kontinents*” in this volume.

She combines methods of distant and close reading and develops new techniques for keyword in context searches that render visible what has not yet been seen in Humboldt's travelogue. In doing so the contribution stresses the necessary combination of both digital and humanities methods in text mining. Peters also addresses the question of our own practices of comparing as humanities researchers and sees new opportunities for these in working with digital methods. She applies this combination of methods not only to test them against her own hypotheses but finds new insights into Humboldt's travel writings along the way.

Anna Dönecke focuses more directly on the question of data modeling in historical research.<sup>46</sup> In her contribution she assumes that data modeling as a basic operation in Digital Humanities can alter the perspective of historians on their sources. Creating a relational database with information from eighteenth-century court records requires a different understanding of their contents, shifting the focus from content information to features and patterns. Her examples show that implementing methods from computer science such as data modeling produces a genuine surplus for historical research. This is especially true when implementing methods of pattern recognition, Dönecke argues, because this explicitly changes the perspective of the researcher towards his or her source material. Using data models and relational databases forces us to dissect the documents we are interested in into tiny bits of information and to attribute meaning to the common features that can be detected by looking at this information rather than by reading the documents as text. It is this way of interacting with textual sources that poses the biggest challenge to our daily work of interpretation as humanities researchers. The contribution by Marcus Hartner, Ralf Schneider, and Anne Lappert demonstrates this nicely.<sup>47</sup> Representing the field of British Literary Studies, the authors went about their project with a clear question in mind. They are interested in the way that the emerging middle class in eighteenth-century Britain represented itself through their morality in contemporary novels. Using Voyant Tools, Hartner et al. look for textual evidence of their hypotheses, but find only little. Their discussion of

---

46 See Dönecke, "From Serial Sources to Modeled Data. Changing Perspectives on Eighteenth-Century Court Records from French Pondicherry" in this volume.

47 See Hartner et al., "Looking for Textual Evidence: Digital Humanities, Middling-Class Morality, and the Eighteenth-Century English Novel" in this volume.



this alleged failure is very enlightening for the relationship of digital and traditional methods. Since their research interest rested with a concept instead of a certain term or phrase in the beginning, the authors test several search strategies to find textual evidence matching their presuppositions. They engage with what has been called “screwmeneutics” diving into the digital tools as a means of explorative hermeneutics.<sup>48</sup>

Digitally enhanced text analysis does not only get more difficult the more complex the task of interpretation is but also the more complex the linguistic structures are that one is looking for. In order to teach sentence structure and the meaning of temporal comparisons to the computer, the tasks of annotating, parsing, and tagging must be applied. The contribution by Willibald Steinmetz, Kirill Postoutenko, Olga Sabelfeld, and Michael Götzelmann discusses the results achieved through tagging in different corpora processed with different taggers, and poses the question whether or not the task of preprocessing is worthwhile when reading and interpreting would do the job at least as fast as the tested taggers did.<sup>49</sup> What takes time is building the models that serve as a basis for (semantic) tagging. And while it is necessary and reasonable to think about the ratio of effort and gain in every project design, the contributions in this volume show that engaging with the digital is worthwhile for the humanities.

## Bibliography

*Anthony, Laurence*, AntConc Homepage, <https://www.laurenceanthony.net/software/antconc/> [accessed: 27.08.2019].

*Anthony, Laurence*, AntConc (Windows, Macintosh OS X, and Linux), <https://www.laurenceanthony.net/software/antconc/releases/AntConc358/help.pdf> [accessed: 27.08.2019].

*Busa, Roberto*, The Annals of Humanities Computing: The Index Thomisticus, in: *Computers and the Humanities* 14 (1980), 83–90.

---

48 The term was coined by *Ramsay, Stephen*, The Hermeneutics of Screwing Around; or What You Do with a Million Books, in: Kevin Kee (ed.), *Pastplay: Teaching and Learning History with Technology*, Ann Arbor: University of Michigan Press, 2014 [2010].

49 See Steinmetz et al., “The Historical Semantics of Temporal Comparisons Through the Lens of Digital Humanities: Promises and Pitfalls” in this volume.

- Epple, Angelika/Erhart, Walter*, Die Welt beobachten – Praktiken des Vergleichens, in: Angelika Epple/Walter Erhart (eds.), *Die Welt beobachten – Praktiken des Vergleichens*, Frankfurt/New York: Campus, 2015, 7–31.
- Heintz, Bettina/Huber, Jörg*, Der verführerische Blick: Formen und Folgen wissenschaftlicher Visualisierungsstrategien, in: Bettina Heintz/Jörg Huber (eds.), *Mit dem Auge denken: Strategien der Sichtbarmachung in wissenschaftlichen und virtuellen Welten (Theorie:Gestaltung 01)*, Zürich/Wien/New York: Voldemeer; Springer, 2001, 9–40.
- Heppler, Jason*, What Is Digital Humanities, <https://whatisdigitalhumanities.com/> [accessed: 21.08.2019].
- Jessop, Martyn*, Digital Visualization as a Scholarly Activity, in: *Literary and Linguistic Computing* 23 (2008), 281–293. doi:10.1093/lc/fqn016.
- McCarty, Willard*, *Humanities Computing*, Houndmills: Palgrave Macmillan, 2014.
- Neubert, Anna/Schwandt, Silke*, Comparing in the Digital Age. The Transformation of Practices, in: Angelika Epple/Walter Erhart/Johannes Grave (eds.): *Practices of Comparing. Towards a New Understanding of a Fundamental Human Practice*. Bielefeld 2020 [in print].
- Ramsay, Stephen*, The Hermeneutics of Screwing Around; or What You Do with a Million Books, in: Kevin Kee (ed.), *Pastplay: Teaching and Learning History with Technology*, Ann Arbor: University of Michigan Press, 2014 [2010].
- Rockwell, Geoffrey/Sinclair, Stéfan*, Text Analysis and Visualization: Making Meaning Count, in: Susan Schreibman/Raymond G. Siemens/John Unsworth (eds.), *A New Companion to Digital Humanities*, Chichester, West Sussex, UK/Boston, Massachusetts: Wiley/Blackwell; Credo Reference, 2018, 274–290.
- Rockwell, Geoffrey/Sinclair, Stéfan*, Voyant. See through your Text, <https://voyant-tools.org/> [accessed: 27.08.2019].
- Rockwell, Geoffrey/Sinclair, Stéfan*, Tools, <https://voyant-tools.org/docs/#!/guide/tools> [accessed: 27.08.2019].
- Schwandt, Silke*, Digitale Objektivität in Der Geschichtswissenschaft? Oder: Kann Man Finden, Was Man Nicht Sucht?, in: *Rechtsgeschichte – Legal History* 24 (2016), 337–338. doi:10.12946/rg24/337-338.
- Schwandt, Silke*, Digitale Methoden Für Die Historische Semantik.: Auf Den Spuren Von Begriffen in Digitalen Korpora, in: *Geschichte und Gesellschaft* 44 (2018), 107–134.

- Terras, Melissa M*, Peering Inside the Big Tent, in: Melissa M. Terras/Julianne Nyhan/Edward Vanhoutte (eds.), *Defining Digital Humanities: A Reader*, London: Routledge, 2016, 263–270.
- Terras, Melissa M./Nyhan, Julianne/Vanhoutte, Edward* (eds.) *Defining Digital Humanities: A Reader*. London: Routledge, 2016.
- Universität Bielefeld*, SFB 1288, Practices of Comparing. Ordering and Changing the World, [https://www.uni-bielefeld.de/\(en\)/sfb1288/](https://www.uni-bielefeld.de/(en)/sfb1288/) [accessed: 21.08.2019].
- Universität Bielefeld*, SFB 1288, TP INF, Data Infrastructure and Digital Humanities, [https://www.uni-bielefeld.de/\(en\)/sfb1288/projekte/inf.html](https://www.uni-bielefeld.de/(en)/sfb1288/projekte/inf.html) [accessed: 21.08.2019].
- Vanhoutte, Edward*, The Gates of Hell: History and Definition of Digital | Humanities | Computing, in: Melissa M. Terras/Julianne Nyhan/Edward Vanhoutte (eds.), *Defining Digital Humanities: A Reader*, London: Routledge, 2016, 119–156.

# **I. Challenges for the Humanities**



# Open Access, Open Data, Open Software?

## Proprietary Tools and Their Restrictions

---

*Helene Schlicht*

### Open Science and the (Digital) Humanities

The goal of this article is to popularize Open Science principles and shed light on the role of Open Science in the research landscape of the humanities in general and Digital Humanities (DH) in particular. The commitment to Open Science is widespread among digital humanists but has not yet gained a similar foothold in the research culture of the humanities in general. Despite there being a lot of proprietary solutions offered for scholars conducting research with the aid of digital methods in the humanities, many digital humanists deem it important to choose only open formats to ensure as much inclusivity as possible (on various levels). It is my intention to make an argument for pushing the implementation of Open Science principles in the humanities and explain why it is crucial – even if it makes work more difficult sometimes. As Siemens suggests, I want to explore “the digital humanities’ positive role in the process of the humanities’ digital self-determination in the digital realm.”<sup>1</sup>

At first the topics discussed in this article may seem to be disparate but I aim to show how they are interwoven and can benefit from and stimulate each other. Open Science principles, if taken seriously, determine the priorities in tool development and usage. As a result, aspects have to be taken into account that would otherwise not have been considered, and a different prioritization of tools and programs needs to come

---

<sup>1</sup> *Siemens, Ray*, Communities of practice, the methodological commons, and digital self-determination in the Humanities., in: *Digital Studies/Le champ numérique* (2016). <http://doi.org/10.16995/dscn.31>.

to effect.<sup>2</sup> However, the implementation of Open Science principles does not happen without stakeholders who actively advance and enforce them – the implementation of these principles largely depends on the relevance attributed to them by the respective fields and researchers. In my contribution, I want to illustrate how the frictions between Digital Humanities and the broader humanities can be utilized to come to a mutual understanding about the implementation of Open Science principles.

To discuss the topic of Open Science, I will draw on concepts of Open Access because this part of Open Science has already been widely discussed, and from there on broaden the subject to other elements of Open Science. Subsequently, I will link this to the relationship of the humanities and Digital Humanities and its potential for extending the practices of humanities research. Here I also want to point out the noteworthy role of funding agencies and universities in driving this development forward.

## Why do we need Digital Humanities?

There is an ongoing discussion about what Digital Humanities is or should be. For this article I will operate with a minimal definition of Digital Humanities and beyond that only address those aspects of Digital Humanities that explain why so many digital humanists emphasize the importance of Open Science. “Digital humanities is a diverse and still emerging field that encompasses the practice of humanities research in and through information technology, and the exploration of how the humanities may evolve through their engagement with technology, media, and computational methods.”<sup>3</sup> Although Digital Humanities is still a part of the humanities which is regarded with some suspicion and sometimes only understood as a service provider for research and for the application of tools, as Sahle remarks<sup>4</sup>, its

---

2 This can for example mean to weigh inclusivity and functionality against each other to negotiate if some cutbacks in seamless functioning are worth the enhanced inclusivity and access for other researchers.

3 *Svensson, Patrik*, The Landscape of Digital Humanities, in: *Digital Humanities Quarterly* 4 (2010). <http://digitalhumanities.org:8081/dhq/vol/4/1/000080/000080.html>.

4 *Sahle, Patrick*, Digital Humanities? Gibt's doch gar nicht!, in: Constanze Baum/Thomas Stäcker (eds.), *Grenzen und Möglichkeiten der Digital Humanities* (= Sonderband der Zeitschrift für digitale Geschichtswissenschaften, 1), 2015. [https://doi.org/10.17175/sb001\\_004](https://doi.org/10.17175/sb001_004).

resources can be used in a productive way to broaden the methodological (and intellectual) framework of the humanities as a whole.

Sahle points out that “[t]he Digital Humanities [...] are embedded in an extensive infrastructure in regard to organization, information, and communication and build upon long traditions in various areas of research. Furthermore, as a link between the humanities and computer science, the field seems to be highly attractive, not only to these areas, but also to neighboring disciplines as well as to the research funding agencies.”<sup>5</sup>

The problem remains that “[w]hen we do try to define [digital humanities] in a way that can lead to action, especially at a local level within an institutional structure, we tend to arrive at institutional- or discipline-specific definitions; these do have some sort of gain in that you can frame digital humanities in the terms of extant structures, but ultimately there’s a loss via disciplinarity’s constraint in light of current and future growth, narrowing potential collaborative opportunities and limiting the vision of what the intersection points between the humanities and digital could lead to.”<sup>6</sup> Instead of focusing on disciplinary boundaries I want to direct the focus on a different aspect. Siemens shifts the discussion from questioning Digital Humanities to asking about the role of the humanities at large in the digital age: “How do the humanities fit in a digital age, reflecting and engaging not only its own traditions but, further, those of other disciplines implicated in, drawn in, partnered with, and fully incorporated and embraced by the methods utilized by the digital humanities. Does it do so by situating itself outside the humanities, outside of the very context that makes digital *humanities* different from other computational enterprises? I’d think not; I’d think we’d ideally work to situate it well within the humanities.”<sup>7</sup>

This changes the focus inasmuch as it implicitly asserts the necessity for the humanities at large to adjust to changing general conditions of doing research. To attune to the digital age and the changes it inevitably brings, the humanities should make use of the shared practices tested over generations and find a way to best transpose them into the digital realm. Digital Humanities can be of discipline-specific as well as infrastructural aid in the process of translation. One way to do so could be to look at the methodolog-

---

5 Ibid.

6 R. Siemens, *Communities of Practice*.

7 Ibid.



ical commonalities of the (digital) humanities in a sense that “[...] the notion of the community of practice here offers us a framework to consider and understand *who* we are via *what* it is we do, *where* we do what we do, and *why* we do it in the way that we do it. What is most unique about this frame is how it focuses us on the set of practices we share, who we share the practices with and where, on what we apply them, and to what end we do so.”<sup>8</sup>

This is crucial because it helps us reflect on what would be essential in developing new (software) tools and solutions to encompass shared practices of humanistic inquiry. It is important to keep in mind that software is not neutral, that the digitization is not neutral and that research cannot be transposed into the digital realm without repercussions we need to reflect on. “[E]ach stage in the digitization [...] has, among other things, semiotic, social, cultural and political implications.”<sup>9</sup> As researchers in the humanities we need to reflect on those implications of the digitization from different vantage points but also keep our own desiderata in mind. We need to know the requirements of working with digital tools to figure out how to implement them into digital technologies. “If we are interested in creating in our work with digital technologies the subjective, inflected, and annotated processes central to humanistic inquiry, we must be committed to designing the digital systems and tools for our future work. Nothing less than the way we understand knowledge and our tasks as scholars are at stake. Software and hardware only put into effect the models structured into their design.”<sup>10</sup> As Drucker describes, trying to put processes of humanistic research into practice in tool design requires the participation of those working in the fields of the humanities. Furthermore, trying to integrate Open Science principles into tool development and implementation brings about changes in emphasis. Different aspects of tools become important. The consideration of Open Science principles thereby leads to questions regarding fairness and inclusivity. I will discuss the entanglement of these topics, but first I need to lay the groundwork for this discussion by elaborating on the state of Open Science.

---

8 Ibid.

9 *Fiormonte, Domenico/Numerico, Teresa/Tomasi, Francesca (eds.)*, *The Digital Humanist. A Critical Inquiry*, trans. by Desmond Schmidt with Christopher Ferguson, New York: Punctum Books 2015, 17.

10 *Drucker, Johanna*, *Blind Spots: Humanists must plan their digital future*, in: *The Chronicle of Higher Education* 55 (2009), B6-B8. <https://www.chronicle.com/article/Blind-Spots/9348>. [Paywalled]

## Open Science, Open Access, Open Data

“Open Science is an umbrella term encompassing a multitude of assumptions about the future of knowledge creation and dissemination”, as Fecher and Friesecke point out.<sup>11</sup> As such, the term Open Science merges a diverse set of ideas and initiatives: Open Access and Open Data Initiatives, Open Scholarship or the demand for Open Educational Resources, Open Source Software, Open Review, Open Metrics, and demands for Open Methodology all get subsumed under Open Science. To define the idea behind Open Science more specifically, the definition of the Open Knowledge Foundation proves helpful and can serve as a minimal consensus: “Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness).”<sup>12</sup> How this minimal consensus is put into practice in individual cases, differs quite drastically. The furthest implementation and greatest acceptance can be attributed to Open Access (OA). Therefore, I use OA as an example to point out the relevance of Open Science principles for the humanities as well as for our specific situation in a collaborative research center. Among the many positive features of Open Access are, for example, a higher visibility, free access for every user – regardless of the researchers’ affiliation –, better retrievability, and a faster dissemination of research results.<sup>13</sup> The OA movement is described as having “[...] two different, alternative, converging histories: the history of the economics of recent academic journal publishing and the history of the free culture movement, which has its roots in the world of computer software.”<sup>14</sup>

---

11 Fecher, *Benedikt/Friesike, Sascha*, Open Science: One Term, Five Schools of Thought, in: Sönke Bartling/Sascha Friesike (eds.), *Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing*, Cham: Springer Open 2014, 17.

12 *Kleineberg, Michael/Kaden, Ben*, Open Humanities? ExpertInnenmeinungen über Open Access in den Geisteswissenschaften, in: LIBREAS. Library Ideas 32 (2017), <https://libreas.eu/ausgabe32/kleineberg/>.

13 *Arbeitsgruppe Open Access in der Allianz der deutschen Wissenschaftsorganisationen*, *Open Access: Positionen, Prozesse, Perspektiven*, Bonn: Köllen Druck+Verlag GmbH, 2009, 3. <http://doi.org/10.2312/allianz0a.001>.

14 *Eve, Martin Paul*, *Open Access and the Humanities: Contexts, Controversies and the Future*, Cambridge: Cambridge University Press, 2014, 12. <https://doi.org/10.1017/CBO9781316161012>.

This might in part explain why the ideals of Open Science and Open Access are more widespread in the DH community than in the broader field of the humanities at the moment.

The debate about Open Access started in the 1990s, when the first influential commitments to Open Access were formulated by a bunch of initiatives, from the Budapest Open Access Initiative<sup>15</sup> to the Bethesda Statement<sup>16</sup> and the Berlin Declaration<sup>17</sup> – to name just the most influential –, all drafted by different stakeholders but demanding similar policies and practices. Humanities scholars were involved in the formulation of all of these statements and hence have been part of the debate from the start.<sup>18</sup> The Berlin Declaration, for example, was drafted at the end of a conference held by the Max-Planck-Society and the project European Cultural Heritage Online (ECHO) and signed by all big scientific organizations in Germany as well as several universities, academies and other research and cultural institutes.<sup>19</sup>

The Bethesda Statement formulates two criteria Open Access publications have to meet, which can also be found in the almost exact same wording – only minor details are added – in the Berlin Declaration:

“1. The author(s) and copyright holder(s) grant(s) to all users a free, irrevocable, worldwide, perpetual right of access to, and a license to copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship, as well as the right to make small numbers of printed copies for their personal use.

---

15 *Budapest Open Access Initiative*, Budapest Open Access Initiative, <https://www.budapestopenaccessinitiative.org/> [accessed: 17.05.2019].

16 *Suber, Peter et al.*, The Bethesda Statement on Open-Access Publishing, (Jun 20, 2003), <http://legacy.earlham.edu/~peters/fos/bethesda.htm> [accessed: 01.09.2019].

17 *Max-Planck-Gesellschaft*, Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (Oct 22, 2003), <https://openaccess.mpg.de/Berliner-Erklaerung> [accessed: 01.09.2019].

18 *M. P. Eve*, *Open Access and the Humanities*, 24.

19 One of the aims of this conference was to think about web-based research environments and the future of scientific publishing online. The conference announcement is available at: *Max-Planck-Gesellschaft*, Berlin-Konferenzen, <https://openaccess.mpg.de/BerlinOA> [accessed: 09.05.2019].

2. A complete version of the work and all supplemental materials, including a copy of the permission as stated above, in a suitable standard electronic format is deposited immediately upon initial publication in at least one online repository that is supported by an academic institution, scholarly society, government agency, or other well-established organization that seeks to enable open access, unrestricted distribution, interoperability, and long-term archiving [...].<sup>20</sup>

From the beginning OA initiatives demanded, as I want to point out, a change in the practices of all stakeholders involved.

Though there also have been critical voices raising concerns about academic freedom<sup>21</sup> and about different logics of publishing between the sciences and the humanities<sup>22</sup> (some of whom I will reference later), there were a lot less counter-initiatives.<sup>23</sup>

Alongside these first declarations there is a wide array of initiatives to make the case for a wider acceptance of Open Access in all parts of academia. In several countries state-led initiatives build regulatory foundations in various ways to fasten the implementation of Open Access. In Germany, the most important research funding agency, the Deutsche Forschungsgemeinschaft (DFG), expects researchers to publish their work Open Access if it is funded

---

20 P. Suber *et al.*, The Bethesda Statement.

21 The two biggest concerns regarding academic freedom are first that mandatory Open Access leaves no room to discuss what the role of academic labor is and whom its merits are granted to and second that it will restrict the authors' ability to say how, where, and by whom her work could be reused. See: *Columbia, David*, Marxism and Open Access in the Humanities: Turning Academic Labor Against Itself, in: *Workplace: A Journal for Academic Labor* 28 (2016), 74–114, esp. pp. 100–101.; *Anderson, Rick*, Open Access and Academic Freedom (Dec 15, 2015), <https://www.insidehighered.com/views/2015/12/15/mandatory-open-access-publishing-can-impair-academic-freedom-essay> [accessed: 31.08.2019].

22 Rosenzweig argues that scholarly societies play a big role in publishing in the humanities and that mandatory Open Access could threaten the societies and their journals because they need the money coming in through journal subscriptions. *Rosenzweig, Roy*, Should Historical Scholarship Be Free?, in: *Roy Rosenzweig, Clio wired: The Future of the Past in the Digital Age*, New York, NY: Columbia Univ. Press, 2011, 119–120.

23 The Heidelberger Appell is an example for an initiative spawn out of a humanities perspective: *Reuß, Roland*, Heidelberger Appell: Für Publikationsfreiheit und die Wahrung der Urheberrechte (Mar 22, 2009), <http://www.textkritik.de/urheberrecht/appell.pdf> [accessed: 01.09.2019].

by the DFG, a purposive policy was approved in 2006.<sup>24</sup> Similar principles were implemented for the whole of Europe through Horizon 2020 and will be adopted and amplified through its successor program Horizon Europe in line with “cOAlition S” and “Plan S”. Those plans were devised by Science Europe, an association of European research funding organizations,<sup>25</sup> together with the European Commission. Their mission is to accelerate “the transition to full and immediate Open Access to scientific publications”<sup>26</sup> to reach a Europe-wide mandatory implementation of OA for research funded by the EU. The Member States of the EU believe “that free access to all scientific publications from publicly funded research is a moral right of citizens” and in 2016 jointly “committed to achieve this goal by 2020.”<sup>27</sup> “Plan S” and “cOAlition S” mean an intensification of previous OA policies. Whereas under the regulations of Horizon 2020 green OA and even hybrid OA<sup>28</sup> met the requirements on Open Access publication, “Plan S” expects of research funders to “[...] mandate that access to research publications that are generated through research grants that they allocate, must be fully and immediately open and cannot be monetised in any way.”<sup>29</sup> This means that publication in hybrid form does not meet the proposed criteria. In an additional statement it is clarified that pre-prints “will satisfy open access requirements” but that “Article Processing Charges will be eligible for purely open access publishing venues (e. g. not ‘hybrid’ journals).”<sup>30</sup> Horizon Europe also intensifies the

24 *Deutsche Forschungsgemeinschaft*, FAQ: Open Access, (last modified Jan 10, 2017), [https://www.dfg.de/foerderung/faq/open\\_access\\_faq/index.html](https://www.dfg.de/foerderung/faq/open_access_faq/index.html) [accessed: 01.09.2019].

25 *Science Europe*, About Us, <https://www.scienceeurope.org/> [accessed: 17.05.2019].

26 *European Commission*, ‘Plan S’ and ‘cOAlition S’ – Accelerating the transition to full and immediate Open Access to scientific publications (Sep 4, 2018), [https://ec.europa.eu/commission/commissioners/2014-2019/moedas/announcements/plan-s-and-coalition-s-accelerating-transition-full-and-immediate-open-access-scientific\\_en](https://ec.europa.eu/commission/commissioners/2014-2019/moedas/announcements/plan-s-and-coalition-s-accelerating-transition-full-and-immediate-open-access-scientific_en) [accessed: 01.09.2019].

27 *Ibid.*

28 Both terms will be explained on pp. 5–6. In short: Green OA means the permission of self-archiving whereas hybrid OA means publishing OA in an otherwise subscription-based journal.

29 *Science Europe*, Science Without Publication Paywalls. Preamble to: cOAlition S for the Realisation of Full and Immediate Open Access (Sep 2018), <https://www.scienceeurope.org/wp-content/uploads/2018/09/cOAlitionS.pdf> [accessed: 01.09.2019].

30 *European Commission*, Horizon Europe Impact Assessment. Staff Working Document 307, Part 2 of 3 (Jun 7, 2018), 106. [https://ec.europa.eu/info/sites/info/files/swd\\_2018\\_307\\_f1\\_impact\\_assesment\\_en\\_v6\\_p2\\_977548.pdf](https://ec.europa.eu/info/sites/info/files/swd_2018_307_f1_impact_assesment_en_v6_p2_977548.pdf) [accessed: 01.09.2019].

regulations on research data as well as other research related output.<sup>31</sup> Science Europe raises the pressure on researchers and publishers alike by stating that “our collective duty of care is for the science system as a whole, and researchers must realise that they are doing a gross disservice to the institution of science if they continue to report their outcomes in publications that will be locked behind paywalls. We also understand that researchers may be driven to do so by a misdirected reward system which puts emphasis on the wrong indicators (e. g., journal impact factor). We therefore commit to fundamentally revise the incentive and reward system of science [...]”<sup>32</sup> This indicates that the goal of these programs is not only an extensive policy shift but a deep impact on today’s research landscape accompanied by and calling for a change in publication practices.

Furthermore, there are joint initiatives to establish nationwide license agreements with big publishers – the German initiative is called Projekt Deal<sup>33</sup> – to secure access to the whole portfolio of e-journals, especially subscription based journals, with the goal to establish better deals and to pressure big publishers into transitioning to OA and publishing all articles of all participating institutions Open Access.<sup>34</sup>

Widening the scope of Open Access and ensuring a wider implementation of Open Science principles are guidelines that were developed for the treatment of research data. The FAIR Guiding principles for scientific data management, for example, consist of four cornerstones that should be con-

---

31 Ibid.

32 *Science Europe*, Science Without Publication Paywalls.

33 *Projekt Deal*, Über DEAL, <https://www.projekt-deal.de/aktuelles/> [accessed: 17.05.2019].

34 A lot is happening in this field right now. While Projekt DEAL established an agreement with Wiley in January, the University of California canceled its subscription to Elsevier in February after unsuccessful negotiations. As UC California is the largest public university system in America, a big impact was expected. Lastly, around Easter, Norway was the first country to strike a deal with Elsevier that allows access to all Elsevier publications and OA publication for Norwegian researchers for a two-year pilot phase. (See: *Projekt DEAL*, Veröffentlichung des Deal-Wiley Agreements, <https://www.projekt-deal.de/vertraagsveroeffentlichung-des-deal-wiley-agreements/> [accessed: 17.05.2019]. *Gaind, Nisha*, Huge US university cancels subscription with Elsevier, in: *Nature* 567 (2019), 15–16. <https://www.nature.com/articles/d41586-019-00758-x>. *Elsevier*, Norway and Elsevier agree on pilot national license for research access and publishing (Apr 23, 2019), <https://www.elsevier.com/about/press-releases/corporate/norway-and-elsevier-agree-on-pilot-national-licence-for-research-access-and-publishing> [accessed: 01.09.2019].

sidered: the organization of data should be executed in a way that data is Findable, Accessible, Interoperable, and Reusable.<sup>35</sup> These principles for data management are not crafted in a way to focus solely on human scholars but instead “[...] put specific emphasis on enhancing the ability of machines to automatically find and use the data [...]” and with the intention that “[...] the principles apply not only to ‘data’ in the conventional sense, but also to the algorithms, tools, and workflows that led to that data.”<sup>36</sup> Guidelines on how to store research data and make it accessible increasingly become part of regulatory efforts.

Other initiatives, like DORA, the San Francisco Declaration on Research Assessment, focus on other parts of Open Science – in this case on Open Metrics. The goal of DORA is to “improve the ways in which the output of scientific research is evaluated by funding agencies, academic institutions, and other parties.”<sup>37</sup> Their goal is to establish new forms of evaluating the quality of research output as alternatives to the flawed Journal Impact Factor.

For more information on the benefits of Free and Open Source Software (FOSS) see the article “From text to Data.”

Before I go on to explain different variants of OA, I need to elaborate further on the term “open” as such. Although I am a proponent of Open Science, I want to make clear that the notion of “open” that is taken up in Open Science principles is in no way unambiguous. As I mentioned earlier, the concept has its roots in movements evolving around computer culture and the free software movement. And this is where its ambiguity stems from. As Evgeny Morozov points out in an extensive essay, “[f]ew words in the English language pack as much ambiguity and sexiness as ‘open.’”<sup>38</sup> He goes on to elaborate that it was a process of active rebranding that led the free software movement to shift from “free” to “open”. “Profiting from the term’s ambiguity, O’Reilly and his collaborators likened the ‘openness’ of open source

---

35 *Wilkinson, Mark et al.*, The FAIR Guiding Principles for scientific data management and stewardship, in: *Scientific Data* 3 (2016): 1-9. <https://doi.org/10.1038/sdata.2016.18>.

36 *Ibid.*

37 *SFDORA*, San Francisco Declaration on Research Assessment, <https://sfdora.org/read/> [accessed: 20.05.2019].

38 *Morozov, Evgeny*, The Meme Hustler: Tim O’Reilly’s crazy talk, in: *The Baffler* 22 (2013), <http://thebaffler.com/salvos/the-meme-hustler> [accessed: 01.09.2019].

software to the ‘openness’ of the academic enterprise, markets, and free speech. ‘Open’ thus could mean virtually anything, from ‘open to intellectual exchange’ [...] to ‘open to competition’.”<sup>39</sup> This ambiguity also shows in the concept of and discussions around Open Access, making it difficult to fully embrace this concept without falling for its ambiguities.

## Variants of Open Access

I want to elaborate very briefly on the different forms of Open Access. One differentiation is made between the so-called green OA and gold OA Standards. Green OA means that research is published in a subscription journal but the researcher retains the right to publish their work in a repository as a pre-print or after a set embargo period, whereas publications under gold standard are published OA right away. The costs of OA publications are normally shifted towards the side of the author in the form of article processing charges (APC)<sup>40</sup> – if APCs are charged at all. This can mean that the author has to pay for making OA available, but normally the cost for publication is eligible for funding – which is mostly explicitly stated by funding agencies.<sup>41</sup> Journals can also decide to waive article fees if the author does not have funding.<sup>42</sup> Another form of OA is hybrid OA, which means that research is published Open Access but in a subscription journal. This practice is highly controversial because it means that article processing charges are imposed for publishing OA while at the same time subscription fees are levied, a practice referred to as “double dipping”. Hybrid Journals are therefore not listed in the Directory of Open Access Journals (DOAJ), a directory indexing peer reviewed Open Access research journals and their metadata.<sup>43</sup> Introduced in 2013, diamond OA is a relatively recent form of OA publish-

---

39 Ibid.

40 The alternative term for this, “author processing fees” and variants thereof are misleading as Suber points out because the fees are rarely paid by the author herself since they are eligible for funding. P. Suber *et al.*, The Bethesda Statement, 138.

41 E. g., *Science Europe*, 10 Principles, <https://www.coalition-s.org/10-principles/> [accessed: 20.05.2019].

42 M. P. Eve, *Open Access and the Humanities*, 59.

43 *Directory of Open Access Journals*, FAQ: What is DOAJ, <https://doaj.org/faq#definition> [accessed: 20.05.2019].



ing.<sup>44</sup> It was defined as a reaction to the trend of gold OA becoming a business model, which the authors trace back to the taking over of the distinction of gold and green through Horizon 2020s research funding program.<sup>45</sup> They fear that this “[...] broad definition of gold OAs ideologically disguises the differences between for-profit and non-profit models and invites ideological abuse of this category by for-profit publishers [...]” which will in turn foster predatory Open Access Journals.<sup>46</sup> In contrast, “[i]n the *Diamond Open Access Model, not-for-profit, non-commercial organizations, associations or networks publish material that is made available online in digital format, is free of charge for readers and authors and does not allow commercial and for-profit re-use.*”<sup>47</sup> Using statistics provided by DOAJ, Fuchs and Sandoval point out that “[...] in September 2013, out of a total of 9891 journals listed in the DOAJ, 6527 (66.0%) had no article processing charges [...]” – with an especially low rate of APC-based journals in the humanities (between 2.3% in History and 28.1% in Business and Management).<sup>48</sup>

All aforementioned variations of OA deal with venues of distribution. Another differentiation that primarily affects the user’s rights or freedom is made between gratis and libre OA. Gratis in this case means the removal of price barriers alone, while libre OA is defined as removing price barriers and a varying range of permission barriers. Suber transposed the terms gratis and libre from software development, where they are used to express the same distinction.<sup>49</sup> Both green and gold OA can be gratis as well as libre, but to obtain libre OA is usually easier for gold OA publications,<sup>50</sup> while diamond-OA is automatically libre. Authors who want to publish their work libre OA need to waive some of their copyrights. This is well regulated through open licenses, e. g., the Creative Commons licenses, which even allow for different gradients of usage approval, while publishing in subscription

---

44 Fuchs, Christian/Sandoval, Marisol, The Diamond Model of Open Access Publishing: Why Policy Makers, Scholars, Universities, Libraries, Labour Unions and the Publishing World Need to Take Non-Commercial, Non-Profit Open Access Serious, in: triple(C) 13 (2013), 428–443.

45 Ibid., 433.

46 Ibid., 436.

47 Ibid., 438.

48 Ibid., 434.

49 P. Suber et al., The Bethesda Statement, 65–66.

50 Ibid., 67.

journals mostly means completely waiving the copyrights of a publication in favor of the publisher. Creative Commons licensing “[...] leaves authors with permanent ownership of their work and lets them reprint that work without seeking permission or paying fees.”<sup>51</sup> The different versions of Creative Commons offer different versions of rights clearance: “Creative Commons offers CC0 (CC-Zero) for copyright holders who want to assign their work to the public domain. The CC Attribution license (CC-BY) describes the least restrictive sort of libre OA after the public domain. It allows any use, provided the user attributes the work to the original author.”<sup>52</sup> Then there are different versions of restricting CC-BY licenses: using CC-BY-NC<sup>53</sup> forbids commercial usage while CC-BY-ND<sup>54</sup> restricts editing, to name just the most known. The regulations made through “Plan S” make some kind of open licensing, preferably CC-BY mandatory.<sup>55</sup> This type of license has been criticized for restricting academic freedom because it “effectively assigns all of the exclusive prerogatives of the copyright holder to the general public, allowing anyone who so desires to copy, distribute, translate, create derivative works, etc., even for commercial purposes, as long as the author is given credit as creator of the original work”,<sup>56</sup> which is especially relevant in the humanities. Critics have also pointed out that it is “[...] wise to be cautious of the fact that the motivation of many governments pursuing open access is to allow industry to take the fruits of (often public) scientific research and to re-enclose it for commercial benefit.”<sup>57</sup> As I already mentioned, “Plan S” also forbids the monetization of research, but this policy only regulates the researcher. Still, these arguments can not be attributed

---

51 *Anderson, Talea/Squires, David*, Open Access and the Theological Imagination, in: *Digital Humanities Quarterly* 11 (2017), <http://www.digitalhumanities.org/dhq/vol/11/4/000340/000340.html>.

52 *P. Suber et al.*, The Bethesda Statement, 69.

53 For more information check: *Creative Commons*, Attribution-NonCommercial 3.0 Germany (CC BY-NC 3.0 DE), <https://creativecommons.org/licenses/by-nc/3.0/de/deed.en> [accessed: 20.05.2019].

54 For more information check: *Creative Commons*, Attribution-NoDerivs 3.0 Germany (CC BY-ND 3.0 DE), <https://creativecommons.org/licenses/by-nd/3.0/de/deed.en> [accessed: 20.05.2019].

55 *Science Europe*, 10 Principles.

56 *R. Anderson*, Open Access and Academic Freedom.

57 *M. P. Eve*, Open Access and the Humanities, 23.

to the licensing itself and it is important to note that open licenses explicitly allow data mining and other forms of digital analysis while traditional publication forms normally do not – which is important for everyone wanting to conduct digital research.

## Why Open Access?

Discussions on Open Access have gained momentum because of the so called “Journal crisis”, which worsened over the last 30 years. The subscription fees of natural science journals have risen and continue to rise because of the formation of large publishing houses through mergers, which led to a quasi-monopolization. At the same time, the measurement of the quality and importance of a journal through impact factors prevailed. A high impact factor means high reputation for the journal but also for every scientist publishing in it because “[t]he JIF, which measures journals’ average citations per article, is often incorrectly used to assess the impact of individual articles.”<sup>58</sup> It mostly also implies high publication and subscription fees.

This also had an impact on the humanities because the risen subscription fees left libraries with less money to buy monographs.<sup>59</sup> The severity of the problem became obvious for the broader public when Harvard’s Faculty Advisory Council signaled that the university’s library could no longer afford the rising cost of subscription fees. The council reported a price increase of about 145% over the past six years, leading them to encourage the “[...] faculty to consider open access publishing as one means of alleviating the high cost of journal subscriptions.”<sup>60</sup> Anderson and Squires point to this as a key moment in the debate because firstly Harvard has the biggest budget of all American universities and secondly it generated publicity.<sup>61</sup> The overall effect of the journal crisis is shown in various studies based on statistics from the

---

58 *Priem, Jason*, *Altmetrics: A manifesto* (October 26, 2010), <http://altmetrics.org/manifesto/> [accessed: 01.09.2019].

59 *Hagner, Michael*, #Open Access: Wie der akademische Kapitalismus die Wissenschaften verändert, in: *Geschichte der Gegenwart* (Sep 26, 2016), [https://geschichtedergegenwart.ch/open\\_access-wie-der-akademische-kapitalismus-die-wissenschaften-veraendert/](https://geschichtedergegenwart.ch/open_access-wie-der-akademische-kapitalismus-die-wissenschaften-veraendert/).

60 *T. Anderson*, *Open Access and the Theological Imagination*.

61 *Ibid.*

Association of Research Libraries. They show that subscription costs “[...] outstripped inflation by over 300 % since 1986.”<sup>62</sup>

The effects on the humanities were profound: “From 1986 to 1997, the unit cost of serials rose 169 percent compared with 62 percent for book monographs. Research libraries’ expenditures for serials thus rose 142 percent compared with only 30 percent for monographs. In 1986 these libraries spent 44 percent of their budgets on books and 56 percent on journals; by 1997 the imbalance had grown to 28 percent for books and 72 percent for journals.”<sup>63</sup> Thus, the journal crisis was followed by a monograph crisis that is mostly felt in the humanities<sup>64</sup> because of the role monographs play in its research culture.

## Problems of inclusivity

Another problem that arises for libraries when they subscribe to journals is that when they “[...] pay for subscriptions to digital journals, they don’t buy or own their own digital copies but merely rent or license them for a period of time. If they cancel a subscription, they could lose access to past issues. They could violate the publishers’ copyrights if they make or hold copies for long-term preservation without special permission or payment [...]”<sup>65</sup> This forces the libraries to carefully negotiate what is at stake in each individual case. I point that out because I see a similar model of dependency growing in the world of software that does not seem to be widely discussed until now. The subscription model of licensing is on the rise in software as well. The fee of software relevant for research processes is paid by the universities, and students and staff can use the programs. Software like *Citavi* has operated on this model for years but it also became more common for other programs formerly using models of perpetual licensing like the *Microsoft Office Suite* or *InDesign*. Also, a lot of proprietary programs in Digital Humanities or digi-

---

62 M. P. Eve, *Open Access and the Humanities*, 13.

63 McPherson, James M., *A Crisis in Scholarly Publishing*, in: *Perspectives on History* 57 (Oct 2003), <https://www.historians.org/publications-and-directories/perspectives-on-history/october-2003/a-crisis-in-scholarly-publishing>.

64 P. Suber et al., *The Bethesda Statement*, 33.

65 Ibid., 34.

tal social sciences operate on equivalent modes. This leads to new forms of exclusivity and exclusion not yet widely problematized by important stakeholders like universities or funding agencies. This university-wide licensing is normally limited to people having at least some affiliation with said university. Thus, universities demand Open Access while at the same time aggravating the problem of access through software licenses. The arguments raised for Open Access are also relevant for this case. Students and researchers get trained in using certain software they might no longer have access to once they leave university. Then they must choose between buying said software or searching for open source alternatives and learning anew how to use them. Training on open source software is still rarely provided at universities. The awareness of the importance of implementing FOSS at university level is still not fully developed but noticeably on the rise – mostly due to questions surrounding data security, protection and sovereignty. But although this problem should be tackled through open science policies and raising awareness, the focus, even of the DH community, seems to lie primarily on workflows instead of the broader implications: “More than causing personal frustration, this reliance on proprietary tools and formats has long-term negative implications for the academic community. In such an environment, journals must outsource typesetting, alienating authors from the material contexts of publication and adding further unnecessary barriers to the unfettered circulation of knowledge.”<sup>66</sup> The argument here is that the reliance on proprietary solutions for scholarly production makes it necessary to outsource parts of the publication process because the scholars cannot do them on their own. I would add that the problem is less the alienation than the business interests of other parties involved (as in the case of the journal crisis). Gil and Ortega add that “[t]he culture of ‘user friendly’ interfaces that has helped popularize computers for almost three decades now, and which underlines the dominant role of .docx, .pdf, and .epub files today, has also led to some basic misunderstandings of what computers can and should do. In the case of writing, the expectation that you should get what you see continues to distance producers from their tools. As with any human tool, we need to understand computers a bit more intimately if we’re going to use

---

66 Tenen, Dennis/Wythoff, Grant, Sustainable Authorship in Plain Text using Pandoc and Markdown, *The Programming Historian* (Mar 19, 2014), <https://doi.org/10.5281/zenodo.1477854>.

them with any degree of critical awareness [...]. [W]hat has remained invisible or grossly misunderstood to producers of scholarship in certain parts of the world are the material conditions of their own knowledge production – digital and analog – with noxious effects for labor and ecological practices.”<sup>67</sup> This argument points to the heart of the problem: Proprietary programs raise the expectation that software should do most of the work themselves and as seamless as possible. Compatibility problems that still often arise when working with FOSS solutions are attributed to the Open Source solution because all worked fine when not using open source – not considering that proprietary programs have a commercial interest in sustaining incompatibilities. The dependencies created by proprietary programs are linked to exclusion mechanisms created through incompatibilities. The problem of inclusivity is still only marginally addressed in the Digital Humanities but is slowly being popularized by DH initiatives like the minimal computing group of `go::dh`,<sup>68</sup> `transform DH`<sup>69</sup> or `Micro DH`.<sup>70</sup>

It remains necessary to continuously raise awareness for the importance of the work that is done in tool development by on the one hand pointing out the restrictions proprietary programs impose on researchers – or sometimes even false assumptions about those programs<sup>71</sup> – and on the other hand disclosing that there are issues of academic freedom and inclusivity involved in making oneself dependent on proprietary programs. Fiormonte,

---

67 Gil, Alex/Ortega, Élika, Global outlooks in digital humanities: Multilingual practices and minimal computing, in: Constance Crompton/Richard J. Lane/Ray Siemens, *Doing Digital Humanities: Practice, Training, Research*, London/New York: Routledge 2016, 30.

68 `GO::DH Minimal Computing working group`, About: What is Minimal Computing, <http://go-dh.github.io/mincomp/about/> [accessed: 20.05.2019].

69 `#transform DH`, About `#transform DH`, <https://transformdh.org/about-transformdh/> [accessed: 20.05.2019].

70 Risam, Roopika/Edwards, Susan, `Micro DH: Digital Humanities at the Small Scale`, in: *Digital Humanities 2017 (2017)*, <http://works.bepress.com/roopika-risam/27/> [accessed: 01.09.2019].

71 In a blog article Thomas Lumley for example responds to a twitter comment of an undisclosed poster complaining about R having no warranties. He responds by citing license agreements of popular proprietary solutions that also do not offer warranties relating to user errors, pointing to this assumption as “a clear symptom of not having read licence [sic] agreements for other statistical software.” Lumley, Thomas, *Absolutely no warranty?* (Feb 18, 2019), <https://notstatschat.rbind.io/2019/02/18/absolutely-no-warranty/> [accessed: 01.09.2019].

Numerico, and Tomasi argue “[...] that humanists need to engage in not only the development of online content but also with ethical issues around computing, especially issues around language, search engines, open access and censorship.”<sup>72</sup>

Hence, in our collaboration we decided to use open source software and solutions wherever possible. As a result, we implemented open source software and tools whenever possible, for other workflows the conversion to open source would be too complicated at the moment because it would produce compatibility issues in a research environment that still mainly uses the Windows operating system. On the other hand, working with Linux based operating systems is in no way trivial. A lot of processes we are by now used to being automated in Windows have to be performed manually, which requires more technological expertise of all parties involved and cannot be implemented without additional training.

A lot needs to be done to raise awareness for the role of the individual researcher as well as universities as driving forces in either perpetuating the dependencies being fostered by using proprietary software or overcoming them. Regarding the implementation of Open Source Software we are still at the beginning. The Open Access movement has a pioneering role now but hopefully paves the way to generate acceptance for the necessary changes in research practices – even if it is more difficult at first. There is no denying that Open Source Software still does not run as smoothly as proprietary programs, but the example of Open Access shows how joint initiatives of important stakeholders can not only shift a discussion but also lead to important policy changes and redirections of money, which, in effect, leads to the emergence of new tools and solutions simplifying the process of OA publishing. Similar effects could be achieved by using the same mechanisms in implementing FOSS and other variants of Open Science. It is important to note that “[...] by making intelligent investments

---

72 Now this points to even broader aspects of the discussed problem: While I follow a line of argument that is fitted for a European context, problems of inclusivity and academic freedom have of course broader and much more serious implications that go beyond the scope of my article. A lot of those problems (from censorship to environmental issues to participation to reducing barriers of all sorts to working conditions) are addressed in the DH initiatives mentioned in Footnotes 67, 68, and 69. D. *Fiormonte/T. Numerico/F. Tomasi (eds.)*, *The Digital Humanist.*, X.

in its information infrastructure, the academia could regain some of its autonomy.”<sup>73</sup>

To achieve this, Open Science principles need to gain a bigger foothold in the overall research culture of the humanities. The Digital Humanities are well suited to be of aid here because there the relevance of the different aspects of Open Science shows regularly in daily endeavors. Therefore, I want to point to discussions on Digital Humanities and its role in the overall humanities research landscape to show how the frictions between both can be utilized not only to prepare the humanities for the demands of the digital age but also to use the critical potential of the humanities for overall changes in its research culture.

### **Frictions in research cultures as starting points for policy changes and metareflections**

The research landscape of Digital Humanities differs in some relevant ways from the established research landscape in the humanities. In part this is due to the questions digital humanists are confronted with in their daily endeavors and the different approaches to the research objects prevalent in the Digital Humanities. “Examples of how and why ‘we’ have to play an active role in the design of the scholarly environments of the future abound in the experience of digital humanists – and are more common in the daily experience of scholars trying to perform basic research and writing tasks than many realize. [...] What version of a work should be digitized as representative of a work?”<sup>74</sup> Or translated into our context: What can be digitized without infringing copyrights? What data can be published? What is research data in the humanities? How can it be published? How can the context of the research be made visible – the material used, its enrichment, the methodology, the people working on making its enrichment possible? Questions that are not at the forefront of humanist thinking constitute the daily endeavors of digital humanists and lead to the recognition of friction points in digital humanities research.

---

73 Fecher, Benedikt/Wagner, Cert G., Open access or the re-conquest of autonomy, in: *encore* (2016), 79, <https://www.hiig.de/encore2016>.

74 J. Drucker, *Blind Spots*.



The importance of cooperation in Digital Humanities (and in Open Science Initiatives, for that matter) shows clearly that the transformation of research practices leads to the recognition of new voids and the development or adaptation of new practices. I will illustrate this with three examples.

The metrics of attribution that are used up until now (and are criticized for various reasons<sup>75</sup>) are not suited for assessing cooperative practices. It became important to find new ways of making the different roles involved in DH cooperations visible by, e. g., developing new ways of highlighting and making attributable the work of researchers involved in programming or maintaining digital research environments. Also, different forms of technology-enabled academic outreach prominent in DH like (micro-)blogging are not accounted for. Therefore, the striving for Open Access also includes new ways of impact measurement trying to depict forms of scientific inquiry that are not accounted for until now – and going a lot further than the DORA-principles. Open Access supporters argue “[...] the case for an alternative and faster impact measurement that includes other forms of publication and the social web coverage of a scientific contribution. The general credo is: As the scholarly workflow is [sic] migrates increasingly to the web, formerly hidden uses like reading, bookmarking, sharing, discussing, and rating are leaving traces online and offer a new basis by which to measure scientific impact. The umbrella term for these new impact measurements

---

75 And, as I want to add, valid reasons. I can not fully encapsulate the discussion but I try to give a very brief description of the main points of criticism: The Journal Impact Factor (JIF) that was originally meant to measure the impact of a journal is often used to derive the (presumed) impact of individual articles, which is in itself an invalid practice. This is in turn employed to assess objectifiable criteria that are used to evaluate the employability of individual researchers. This is a criticizable practice and even worsened by being built on misused parameters. While it is positive that because of new means of academic outreach new forms of impact assessment have been created that can supersede the old, flawed forms of impact measurement, this leaves aside the fundamental discussion on the problems of condensing academic work and impact into quantifiable aspects. For more information see: *Callaway, Ewen*, Beat it, impact factor! Publishing elite turns against controversial metric, in: *Nature* 535 (2016), 210–211. <https://doi.org/10.1038/nature.2016.20224>. *Fenner, Martin*, Altmetrics and Other Novel Measures for Scientific Impact, in: *Sönke Bartling/Sascha Friesike* (eds.), *Opening science: The evolving guide on how the internet is changing research, collaboration and scholarly publishing*. Cham: Springer Open, 2014, 179–189. *Lariviere, Vincent/Sugimoto, Cassidy R.*, The Journal Impact Factor: A brief history, critique, and discussion of adverse effects (Jan 26, 2018), <http://arxiv.org/pdf/1801.08992v2>.

is altmetrics.”<sup>76</sup> Altmetrics are also suitable to comply with other forms of Open Science that gained influence in recent years like Open Research Data or other forms of “raw science”. But they are only one part of the so called scientometrics, focusing on the web, interconnected tools, and social media as new sources for impact measurement. “Altmetrics can help researchers demonstrate the impact of their research, in particular if the research outputs are not journal articles, but datasets, software, etc., and if the impact is best demonstrated in metrics other than citations.”<sup>77</sup>

Furthermore, as Niebisch points out, practices integral for software development and project management like agility<sup>78</sup> or versioning are more and more incorporated into the methodology of Digital Humanities because of the role of cooperative practices.<sup>79</sup> He compares the practices used in the development of digital objects to practices prevalent in philology, differentiating them into the role of versioning and the option of continuous development and points to the potential these practices can unfold in the humanities. He argues that the need for continuous development of software leads to the imperative of thorough documentation. This is what enables cooperative work on a project. Also, because software needs to be maintained and improved continuously, different versions of a program emerge over time. So software development does not create a static product but a historized and archived output. This can be compared to practices used in philology because in both cases texts are enriched by certain criteria. But whereas in philology the final product is a finished edition (at least up until now), in software development there is no final product but a continuous need for improvement – hence the need for versioning.<sup>80</sup> And through the change of practices facilitated by the Digital Humanities, these practices will take root in the humanities.

---

76 B. Fecher /S. Friesike, *Open Science: One Term, Five Schools of Thought*, 40.

77 M. Fenner, *Altmetrics and Other Novel Measures for Scientific Impact*, 183.

78 Agile software development is a diverse set of methods and practices developed to handle work in collaborative self-organizing and cross-functional teams.

79 Niebisch, Arndt, *Agilität, Versionierung und Open Source: Softwareentwicklung und Praktiken der Geisteswissenschaften*, in: *Wie Digitalität die Geisteswissenschaften verändert: Neue Forschungsgegenstände und Methoden* (=Sonderband der Zeitschrift für digitale Geisteswissenschaften) 3 (2018). [https://doi.org/10.17175/SB003\\_009](https://doi.org/10.17175/SB003_009). [http://www.zfdg.de/sb003\\_009](http://www.zfdg.de/sb003_009).

80 Ibid.

Digital Humanities products are characterized by a collaborative structure that vehemently differs from the research methodologies in the humanities. This structure can potentially transform practices in the humanities. As Liu suggests “[...] the appropriate, unique contribution that the digital humanities can make to cultural criticism at the present time is to use the tools, paradigms, and concepts of digital technologies to help rethink the idea of instrumentality. [...] The goal is to rethink instrumentality so that it includes both humanistic and stem (science, technology engineering and mathematics) fields in a culturally broad, and not just narrowly purposive, ideal of service.”<sup>81</sup> The humanities should utilize the critical attitude with which they approach their objects of research for a critical self-examination regarding their own methods and results. This would be a great starting point to evaluate which questions should be addressed and what would be important in tool development. Humanists should take a continuous part in tool development because they could ensure that the diverse iterations during the development contribute to advance the implementation of humanistic paradigms of knowledge and inquiry.

As a third example, the problems of attributing credentials especially for software development in the Digital Humanities spawned the development of principles for software citation. Laurence Anthony, whose tool AntConc was used in our collaboration, suggests forms of citation for the tools he develops on their websites.<sup>82</sup> Besides, several other parties suggest a citation format, among these are the APA,<sup>83</sup> the software sustainability institute<sup>84</sup>

---

81 Liu, Alan, Where is cultural criticism in the digital humanities?, in: Matthew K. Gold (ed.), *Debates in the digital humanities*, Minneapolis: University of Minnesota Press, 2012, 501–502.

82 Laurence Anthony suggests the following citation for AntConc: Anthony, L. (YEAR OF RELEASE). AntConc (Version VERSION NUMBER) [Computer Software]. Tokyo, Japan: Waseda University. Available from <http://www.laurenceanthony.net/software>.

83 See *Purdue University Online Writing Lab*, Reference List: Electronic Sources, [https://owl.purdue.edu/owl/research\\_and\\_citation/apa\\_style/apa\\_formatting\\_and\\_style\\_guide/reference\\_list\\_electronic\\_sources.html](https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_formatting_and_style_guide/reference_list_electronic_sources.html) [accessed: 21.05.2019].

84 *Software Sustainability Institute*, How to cite and describe the software you used in your research – top ten tips, by Mike Jackson (Jun 22, 2012), <https://www.software.ac.uk/blog/2016-10-07-how-cite-and-describe-software-you-used-your-research-top-ten-tips>.

and the FORCE11 group, whose principles for software citation were published in 2016.<sup>85</sup> In this instance, the humanities were not influenced by computational methods, but rather – via working on a way to incorporate principles of Open Science into the system of attributing credentials – yielded changes in practices of the computational sciences.

These are just a few examples to illustrate how the frictions between the research cultures of traditional humanities and Digital Humanities can be productive by initiating critical reflections on how the landscape of research could evolve. Liu sums up the point I want to make quite nicely when he states that

“[...] digital technology is on the threshold of making a fundamental difference in the humanities because it indeed serves as the vector that imports alien paradigms of knowledge. In terms of objects of inquiry, it brings into play whole new classes or levels of phenomena – e. g. quantitatively defined structures, forms, and cycles. In terms of analytical procedures, digital technology introduces modeling and other kinds of activities to complement interpretation. And in terms of the output or product of knowledge, digital technology expands the repertory of the monograph, essay, and talk (the staples of the humanities) to include programs, databases, visualizations, graphs, maps, etc.”<sup>86</sup>

After having discussed the principles of Open Science with a specific focus on Open Access, the remainder of this text will focus on the role Digital Humanities could play in disseminating these principles across the broader culture of the humanities. The discourses and questions prevalent in Digital Humanities could be of aid when addressing the questions the humanities have to solve in order to adjust their research practices to the demands of the digital age. I will continue with describing aspects of how we tried to tackle these questions in our collaboration and how collaborative research centers in general can be a great facilitator in this process.

---

85 *Smith, Arfon M./Katz, Daniel S./Niemeyer/Kyle E.*, Software citation principles, in: *PeerJ Computer Science* 2 (2016), e86. <https://doi.org/10.7717/peerj-cs.86>.

86 *Liu, Alan*, Digital Humanities and Academic Change, in: *English Language Notes* 47 (2009), 27.

## Examples for the implementation of Open Science principles

The role of collaborative research centers in the humanities should not be underestimated because “[a]t their best, humanities centers and cross-disciplinary institutes are catalysts for humanities-wide perspectives and change.”<sup>87</sup> Woodward asserts that humanities centers “[...] have served as sites for innovation, as laboratories for incubating emerging modes of knowledge and investigating new objects of study in cross-disciplinary and interdisciplinary contexts.”<sup>88</sup> Liu claims that a big advantage of collaborative research centers is that they evolve intellectually around a shared topic.<sup>89</sup> This opens up new possibilities for discussion. As I already pointed out, one big opportunity of collaborative research centers seems to be the option of reciprocal stimulation of research cultures. To gain mutual understanding it is necessary to make explicit and verbalize aspects of research cultures that are assumed to be self-evident. There is great potential in this.

I do not have to explain much regarding our cooperation because Anna Maria Neubert’s contribution to this publication explains it in depth. I just want to sum up that we are part of a large collaborative research center that deals with questions surrounding aspects of practices of comparing, and our teams is responsible for bringing digital research and data management into the research alliance. The collaborative research center consists of eighteen subprojects. Six of them collaborate closely with us and contributed to this book. The projects were originally not designed to conduct digital research – planning how digital methods could be of aid in their research projects was part of the process of constituting our teamwork.

Before I outline which tools we chose to implement, I want to point out that, as is the case with Open Access, the sustainable implementation of open source software and other principles of Open Science needs the engagement of universities and other important research institutions and funding agencies. In some respects this seems to have gained momentum in recent years. The rectorate of Bielefeld University decided to regulate the

---

87 P. Svensson, *The Landscape of Digital Humanities*.

88 Woodward, *Kathleen*, *The Future of the Humanities – In the Present & in Public*, in: *Daedalus* 138 (2009): 113.

89 A. Liu, *Digital Humanities and Academic Change*, 22.

usage of at least some proprietary software on university computers<sup>90</sup> as well as the usage of cloud storage – especially in connection with sensitive data because of uncertainties in data sovereignty.<sup>91</sup> In addition to that, the federal state of North Rhine-Westphalia was the first in Germany to roll out a cloud storage solution for universities and research institutes called *Sciebo*,<sup>92</sup> which makes the data stored there subject to the German Federal Data Protection Act. A similar Europe-wide initiative is on its way in form of the European Open Science Cloud.<sup>93</sup> The next step forward for *Sciebo* is to extend its use cases by making it the basis for a new integrated solution for research data management and adding features that support collaborative work practices. In January of 2019, the universities of Münster, Bielefeld, and Duisburg-Essen started a joint venture financed by the DFG to achieve just that.<sup>94</sup> Again, there is a similar Europe-wide pilot initiative, the EC Open Research Data Pilot called OpenAIRE. It obliges projects it funds to develop (and keep updated) a Data Management Plan and to provide open access to research data, if possible.<sup>95</sup> Initiatives like that require the realignment of research data management practices and of programs facilitating the research process that are used on a daily basis. As pointed out before, this is something that is not yet conclusively resolved in the humanities. Research data management in the humanities begins with discussing what research data in the humanities could consist of. As part of our collaboration, we conducted workshops dedicated to this question,

---

90 *Universität Bielefeld*, IT-Sicherheitsrichtlinie zur Nutzung von Skype. Version 1.0 (Jun 21, 2012), [http://www.uni-bielefeld.de/informationssicherheit/Regelungen/IT-Sicherheitsrichtlinie\\_Skype\\_2012-06-21.pdf](http://www.uni-bielefeld.de/informationssicherheit/Regelungen/IT-Sicherheitsrichtlinie_Skype_2012-06-21.pdf) [accessed: 01.09.2019].

91 *Universität Bielefeld*, IT-Sicherheitsrichtlinie zur Nutzung von Netzlaufwerken und Cloud-Speicher-Diensten. Version 1.0 (Nov 13, 2015), [http://www.uni-bielefeld.de/informationssicherheit/Regelungen/IT-Sicherheitsrichtlinie\\_Cloud-Speicher\\_2015-11-17.pdf](http://www.uni-bielefeld.de/informationssicherheit/Regelungen/IT-Sicherheitsrichtlinie_Cloud-Speicher_2015-11-17.pdf) [accessed: 01.09.2019].

92 *Sciebo*, Das Projekt, <https://www.sciebo.de/projekt/index.html> [accessed: 13.03.2019].

93 *European Commission*, European Open Science Cloud (EOSC), <https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud> [accessed: 20.05.2019].

94 *Universität Bielefeld*, Blog: uni-intern (8 Jan 8, 2019), [https://ekvv.uni-bielefeld.de/blog/uniintern/entry/geb%C3%Bcndelte\\_expertise\\_f%C3%Bcr\\_effizientes\\_forschungsdaten](https://ekvv.uni-bielefeld.de/blog/uniintern/entry/geb%C3%Bcndelte_expertise_f%C3%Bcr_effizientes_forschungsdaten) [accessed: 01.09.2019].

95 *OpenAIRE*, Factsheet H2020 Open Data Pilot, <https://www.openaire.eu/factsheet-h2020-odp> [accessed: 20.05.2019].

and Silke Schwandt held a keynote lecture on this question at a conference in Paderborn in 2018.<sup>96</sup>

To make the matter even more complex, implementing principles of open data in the humanities is far from easy because most of its research objects are subject to copyright law, which is one of the reasons why Open Access receives such broad support in the DH community. Our solution is far from perfect because different tasks as of yet have to be accomplished on different platforms but it is a starting point we can build on since we are still at the beginning of our collaboration. We wanted to publish as much of the data we enriched as possible so we chose DKAN,<sup>97</sup> a free and open source data platform, to collect and publish the gathered research data. DKAN allows for the management of diverse data sets, which includes different gradients of accessibility rights – important in dealing with research data that cannot be published due to copyrights. So some of the data we uploaded will be open for the public, other data will only be accessible by the researcher working with it.

We decided to use Redmine for project management and documentation for the whole collaborative research center and initially used the already existing platform Sciebo to transfer files for further processing. We implemented a pipeline for digitization and natural language processing that is explained further in the article “From Text to Data.” The enriched data generated in this pipeline was then analyzed with several tools depending on the research questions. All of these tools are explained in the introduction to this volume and in Anna Maria Neubert’s article.

A remaining problem of product development within the contexts of Digital Humanities is that “[t]he user interface for many digital projects often seems developed as an afterthought, thrown together after completing the core functionality. However, a truly good user interface requires significant investment in design and development that needs to be integrated into

---

96 *Schwandt, Silke*, Quellen, Daten, Interpretationen: Heterogene Forschungsdaten und ihre Publikation als Herausforderung in der Geschichtswissenschaft, Paper presented at Forschungsdaten in der Geschichtswissenschaft, Jun, 7-8, 2018, Paderborn University, <https://kw.uni-paderborn.de/historisches-institut/zeitgeschichte/veranstaltungen/tagung-forschungsdaten/>.

97 DKAN, “DKAN Open Data Portal, <https://docs.getdkan.com/en/latest/> [accessed: 20.05.2019].

the project timeline and budget.”<sup>98</sup> It can be argued that until now Digital Humanities has neglected to think about interface design and the linked approachability. McGann points out that “[d]igital instruments are only as good as the interfaces by which we think through them.”<sup>99</sup> One unsolved problem that also showed in our cooperation is that tools become increasingly difficult to handle with growing functionality.

Therefore, humanists and digital humanists should join forces in the process of implementing and developing digital tools because “[t]he task of modeling an environment for scholarship (not just individual projects, but an environment, with a suite of tools for access, use, and research activity) is not a responsibility that can be offloaded onto libraries or technical staffs. [...] *The design of digital tools for scholarship is an intellectual responsibility, not a technical task.*”<sup>100</sup> This is not to be underestimated. “The scope of the task ahead is nothing short of modeling scholarly activity anew in digital media.”<sup>101</sup> But if we are not involved in this process of “designing the working environments of our digital future, we will find ourselves in a future that doesn’t work, without the methods and materials essential to our undertakings.”<sup>102</sup>

A finding that the survey of Gibbs and Owens points to is that humanists would be interested in tools that produce interesting results in a short time – an experience we also made in our cooperation. So perhaps “such rough and ready use should be a more explicit aim of digital humanities tool development. [...] [T]he fundamental barrier to wider adoption of digital tools seems to lie now in quality interfaces, accessible documentation and expectations management.”<sup>103</sup>

We try to take this suggestion seriously for the next step of our collaboration, which will be the implementation of a Virtual Research Environment (all of it Open Source, of course).

---

98 Gibbs, Fred/Owens, Trevor, Building Better Digital Humanities Tools: Toward broader audiences and user-centered designs, in: Digital Humanities Quarterly 6, no. 2 (2012). <http://digitalhumanities.org/8081/dhq/vol/6/2/000136/000136.html>.

99 McGann, Jerome, *The Scholar’s Art: Literary Studies in a Managed World*, Chicago: Chicago University Press, 2006, 156–157. Cited after: Svennson, Patrick, Humanities Computing as Digital Humanities, in: Digital Humanities Quarterly 3 (2009), <http://www.digitalhumanities.org/dhq/vol/3/3/000065/000065.html>.

100 J. Drucker, *Blind Spots*.

101 Ibid.

102 Ibid.

103 F. Gibbs/T. Owens, *Building Better Digital Humanities Tools*.



## Conclusion

Developing, broadening, and popularizing Open Science principles is one of the next big tasks the scientific community must address. The humanities in particular have to find their own tailor-made solutions for the specific requirements of the research processes in the humanities. I have discussed aspects that should be considered in this process of transforming research practices to meet the demands of the digital realm. The digital humanities community has pointed out the specific requirements. Their suggestions could serve as a good starting point for the necessary discussion. As I have shown by taking the example of Open Science, focal points in future tool development must be negotiated. These focal points will in turn determine which questions need to be addressed and which aspects of a tool will be important besides “mere” functionality. This works especially well in cooperative projects because of the option of mutual cross-pollination of research cultures. Furthermore, the discussions facilitated through interdisciplinary cooperations can make the needs of involved research cultures more explicit because they have to be verbalized for the sake of a mutual understanding.

## Bibliography

- #transform DH*, About #transform DH, <https://transformdh.org/about-transformdh/> [accessed: 20.05.2019].
- Anderson, Rick*, Open Access and Academic Freedom (Dec 15, 2015), <https://www.insidehighered.com/views/2015/12/15/mandatory-open-access-publishing-can-impair-academic-freedom-essay> [accessed: 31.08.2019].
- Anderson, Talea/Squires, David*, Open Access and the Theological Imagination, in: *Digital Humanities Quarterly* 11 (2017), <http://www.digitalhumanities.org/dhq/vol/11/4/000340/000340.html>.
- Arbeitsgruppe Open Access in der Allianz der deutschen Wissenschaftsorganisationen*, Open Access: Positionen, Prozesse, Perspektiven, Bonn: Köllen Druck + Verlag GmbH, 2009, <http://doi.org/10.2312/allianz0a.001>.
- Budapest Open Access Initiative*, Budapest Open Access Initiative, <https://www.budapestopenaccessinitiative.org/> [accessed: 17.05.2019].

- Callaway, Ewen*, Beat it, impact factor! Publishing elite turns against controversial metric, in: *Nature* 535 (2016), 210–211. <https://doi.org/10.1038/nature.2016.20224>.
- Creative Commons*, Attribution-NonCommercial 3.0 Germany (CC BY-NC 3.0 DE), <https://creativecommons.org/licenses/by-nc/3.0/de/deed.en> [accessed: 20.05.2019].
- Creative Commons*, Attribution-NoDerivs 3.0 Germany (CC BY-ND 3.0 DE), <https://creativecommons.org/licenses/by-nd/3.0/de/deed.en> [accessed: 20.05.2019].
- Deutsche Forschungsgemeinschaft*, FAQ: Open Access, (last modified Jan 10, 2017), [https://www.dfg.de/foerderung/faq/open\\_access\\_faq/index.html](https://www.dfg.de/foerderung/faq/open_access_faq/index.html) [accessed: 01.09.2019].
- Directory of Open Access Journals*, FAQ: What is DOAJ, <https://doaj.org/faq#definition> [accessed: 20.05.2019].
- DKAN, “DKAN Open Data Portal, <https://docs.getdkan.com/en/latest/> [accessed: 20.05.2019].
- Drucker, Johanna*, Blind Spots: Humanists must plan their digital future, in: *The Chronicle of Higher Education* 55 (2009), B6-B8. <https://www.chronicle.com/article/Blind-Spots/9348>.
- Elsevier*, Norway and Elsevier agree on pilot national license for research access and publishing (Apr 23, 2019), <https://www.elsevier.com/about/press-releases/corporate/norway-and-elsevier-agree-on-pilot-national-licence-for-research-access-and-publishing> [accessed: 01.09.2019].
- European Commission*, Horizon Europe Impact Assessment. Staff Working Document 307, Part 2 of 3 (Jun 7, 2018), [https://ec.europa.eu/info/sites/info/files/swd\\_2018\\_307\\_f1\\_impact\\_assesment\\_en\\_v6\\_p2\\_977548.pdf](https://ec.europa.eu/info/sites/info/files/swd_2018_307_f1_impact_assesment_en_v6_p2_977548.pdf) [accessed: 01.09.2019].
- European Commission*, ‘Plan S’ and ‘cOAlition S’ – Accelerating the transition to full and immediate Open Access to scientific publications (Sep 4, 2018), [https://ec.europa.eu/commission/commissioners/2014-2019/moedas/announcements/plan-s-and-coalition-s-accelerating-transition-full-and-immediate-open-access-scientific\\_en](https://ec.europa.eu/commission/commissioners/2014-2019/moedas/announcements/plan-s-and-coalition-s-accelerating-transition-full-and-immediate-open-access-scientific_en) [accessed: 01.09.2019].
- European Commission*, European Open Science Cloud (EOSC), <https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud> [accessed: 20.05.2019].

- Eve, Martin Paul*, *Open Access and the Humanities: Contexts, Controversies and the Future*, Cambridge: Cambridge University Press, 2014. <https://doi.org/10.1017/CBO9781316161012>.
- Fecher, Benedikt/Wagner, Gert G.*, Open access or the re-conquest of autonomy, in: *encore* (2016), 79, <https://www.hiig.de/encore2016>.
- Fecher, Benedikt/Friesike, Sascha*, Open Science: One Term, Five Schools of Thought, in: Sönke Bartling/Sascha Friesike (eds.), *Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing*, Cham: Springer Open 2014, 17–48.
- Fenner, Martin*, Altmetrics and Other Novel Measures for Scientific Impact, in: Sönke Bartling/Sascha Friesike (eds.), *Opening science: The evolving guide on how the internet is changing research, collaboration and scholarly publishing*, Cham: Springer Open, 2014, 179–189.
- Fiormonte, Domenico/Numerico, Teresa/Tomasi, Francesca (eds.)*, *The Digital Humanist. A Critical Inquiry*, trans. by Desmond Schmidt with Christopher Ferguson, New York: Punctum Books 2015.
- Fuchs, Christian/Sandoval, Marisol*, The Diamond Model of Open Access Publishing: Why Policy Makers, Scholars, Universities, Libraries, Labour Unions and the Publishing World Need to Take Non-Commercial, Non-Profit Open Access Serious, in: *triple(C)* 13 (2013), 428–443.
- Gaind, Nisha*, Huge US university cancels subscription with Elsevier, in: *Nature* 567 (2019), 15–16. <https://www.nature.com/articles/d41586-019-00758-x>.
- Gibbs, Fred/Owens, Trevor*, Building Better Digital Humanities Tools: Toward broader audiences and user-centered designs, in: *Digital Humanities Quarterly* 6, no. 2 (2012). <http://digitalhumanities.org:8081/dhq/vol/6/2/000136/000136.html>.
- Gil, Alex/Ortega, Élika*, Global outlooks in digital humanities: Multilingual practices and minimal computing, in: Constance Crompton/Richard J. Lane/Ray Siemens, *Doing Digital Humanities: Practice, Training, Research*, London/New York: Routledge 2016, 22–34.
- GO::DH Minimal Computing working group*, About: What is Minimal Computing, <http://go-dh.github.io/mincomp/about/> [accessed: 20.05.2019].
- Golumbia, David*, Marxism and Open Access in the Humanities: Turning Academic Labor Against Itself, in: *Workplace: A Journal for Academic Labor* 28 (2016), 74–114.
- Hagner, Michael*, #Open Access: Wie der akademische Kapitalismus die Wissenschaften verändert, in: *Geschichte der Gegenwart* (Sep 26, 2016),

- [https://geschichtedergewandert.ch/open\\_access-wie-der-akademische-kapitalismus-die-wissenschaften-veraendert/](https://geschichtedergewandert.ch/open_access-wie-der-akademische-kapitalismus-die-wissenschaften-veraendert/).
- Kleineberg, Michael/Kaden, Ben*, Open Humanities? ExpertInnenmeinungen über Open Access in den Geisteswissenschaften, in: LIBREAS. Library Ideas 32 (2017), <https://libreas.eu/ausgabe32/kleineberg/>.
- Lariviere, Vincent/Sugimoto, Cassidy R.*, The Journal Impact Factor: A brief history, critique, and discussion of adverse effects (Jan 26, 2018), <http://arxiv.org/pdf/1801.08992v2>.
- Liu, Alan*, Where is cultural criticism in the digital humanities?, in: Matthew K. Gold (ed.), Debates in the digital humanities, Minneapolis: University of Minnesota Press, 2012, 501–502.
- Liu, Alan*, Digital Humanities and Academic Change, in: English Language Notes 47 (2009), 27.
- Lumley, Thomas*, Absolutely no warranty? (Feb 18, 2019), <https://notstatschat.rbind.io/2019/02/18/absolutely-no-warranty/> [accessed: 01.09.2019].
- Max-Planck-Gesellschaft*, Berlin-Konferenzen, <https://openaccess.mpg.de/BerlinOA> [accessed: 09.05.2019].
- Max-Planck-Gesellschaft*, Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (Oct 22, 2003), <https://openaccess.mpg.de/Berliner-Erklaerung> [accessed: 01.09.2019].
- McGann, Jerome*, The Scholar's Art: Literary Studies in a Managed World, Chicago: Chicago University Press, 2006.
- McPherson, James M.*, A Crisis in Scholarly Publishing, in: Perspectives on History 57 (Oct 2003), <https://www.historians.org/publications-and-directories/perspectives-on-history/october-2003/a-crisis-in-scholarly-publishing>.
- Morozov, Evgeny*, The Meme Hustler: Tim O'Reilly's crazy talk, in: The Baffler 22 (2013), <http://thebaffler.com/salvos/the-meme-hustler> [accessed: 01.09.2019].
- Niebsch, Arndt*, Agilität, Versionierung und Open Source: Softwareentwicklung und Praktiken der Geisteswissenschaften, in: Wie Digitalität die Geisteswissenschaften verändert: Neue Forschungsgegenstände und Methoden (= Sonderband der Zeitschrift für digitale Geisteswissenschaften) 3 (2018). [https://doi.org/10.17175/SB003\\_009](https://doi.org/10.17175/SB003_009). [http://www.zfdg.de/sb003\\_009](http://www.zfdg.de/sb003_009).
- OpenAIRE*, Factsheet H2020 Open Data Pilot, <https://www.openaire.eu/factsheet-h2020-odp> [accessed: 20.05.2019].
- Priem, Jason*, Altmetrics: A manifesto (October 26, 2010), <http://altmetrics.org/manifesto/> [accessed: 01.09.2019].

- Projekt Deal*, Über DEAL, <https://www.projekt-deal.de/aktuelles/> [accessed: 17.05.2019].
- Purdue University Online Writing Lab*, Reference List: Electronic Sources, [https://owl.purdue.edu/owl/research\\_and\\_citation/apa\\_style/apa\\_formatting\\_and\\_style\\_guide/reference\\_list\\_electronic\\_sources.html](https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_formatting_and_style_guide/reference_list_electronic_sources.html) [accessed: 21.05.2019].
- Reuß, Roland*, Heidelberger Appell: Für Publikationsfreiheit und die Wahrung der Urheberrechte (Mar 22, 2009), <http://www.textkritik.de/urheberrecht/appell.pdf> [accessed: 01.09.2019].
- Risam, Roopika/Edwards, Susan*, Micro DH: Digital Humanities at the Small Scale, in: *Digital Humanities 2017 (2017)*, <http://works.bepress.com/roopika-risam/27/> [accessed: 01.09.2019].
- Rosenzweig, Roy*, Should Historical Scholarship Be Free?, in: *Roy Rosenzweig, Clio wired: The Future of the Past in the Digital Age*, New York, NY: Columbia Univ. Press, 2011, 117–123.
- Sahle, Patrick*, Digital Humanities? Gibt's doch gar nicht!, in: Constanze Baum/Thomas Stäcker (eds.), *Grenzen und Möglichkeiten der Digital Humanities (= Sonderband der Zeitschrift für digitale Geschichtswissenschaften, 1)*, 2015. [https://doi.org/10.17175/sb001\\_004](https://doi.org/10.17175/sb001_004).
- Sciebo*, Das Projekt, <https://www.sciebo.de/projekt/index.html> [accessed: 13.03.2019].
- Schwandt, Silke*, Quellen, Daten, Interpretationen: Heterogene Forschungsdaten und ihre Publikation als Herausforderung in der Geschichtswissenschaft, Paper presented at *Forschungsdaten in der Geschichtswissenschaft*, Jun, 7-8, 2018, Paderborn University, <https://kw.uni-paderborn.de/historisches-institut/zeitgeschichte/veranstaltungen/tagung-forschungsdaten/>.
- Science Europe*, About Us, <https://www.scienceeurope.org/> [accessed: 17.05.2019].
- Science Europe*, Science Without Publication Paywalls. Preamble to: cOAlition S for the Realisation of Full and Immediate Open Access (Sep 2018), <https://www.scienceeurope.org/wp-content/uploads/2018/09/cOAlitionS.pdf> [accessed: 01.09.2019].
- Science Europe*, 10 Principles, <https://www.coalition-s.org/10-principles/> [accessed: 20.05.2019].
- SFDORA*, San Francisco Declaration on Research Assessment, <https://sfdora.org/read/> [accessed: 20.05.2019].

- Siemens, Ray*, Communities of practice, the methodological commons, and digital self-determination in the Humanities., in: Digital Studies/Le champ numérique (2016). <http://doi.org/10.16995/dscn.31>.
- Smith, Arfon M./Katz, Daniel S./Niemeyer/Kyle E.*, Software citation principles, in: PeerJ Computer Science 2 (2016), e86. <https://doi.org/10.7717/peerj-cs.86>.
- Software Sustainability Institute*, How to cite and describe the software you used in your research – top ten tips, by Mike Jackson (Jun 22, 2012), <https://www.software.ac.uk/blog/2016-10-07-how-cite-and-describe-software-you-used-your-research-top-ten-tips>.
- Suber, Peter et al.*, The Bethesda Statement on Open-Access Publishing, (Jun 20, 2003), <http://legacy.earlham.edu/~peters/fos/bethesda.htm> [accessed: 01.09.2019].
- Svensson, Patrik*, The Landscape of Digital Humanities, in: Digital Humanities Quarterly 4 (2010). <http://digitalhumanities.org:8081/dhq/vol/4/1/000080/000080.html>.
- Svensson, Patrick*, Humanities Computing as Digital Humanities, in: Digital Humanities Quarterly 3 (2009), <http://www.digitalhumanities.org/dhq/vol/3/3/000065/000065.html>.
- Tenen, Dennis/Wythoff, Grant*, Sustainable Authorship in Plain Text using Pandoc and Markdown, The Programming Historian (Mar 19, 2014), <https://doi.org/10.5281/zenodo.1477854>.
- Universität Bielefeld*, IT-Sicherheitsrichtlinie zur Nutzung von Netzlaufwerken und Cloud-Speicher-Diensten. Version 1.0 (Nov 13, 2015), [http://www.uni-bielefeld.de/informationssicherheit/Regelungen/IT-Sicherheitsrichtlinie\\_Cloud-Speicher\\_2015-11-17.pdf](http://www.uni-bielefeld.de/informationssicherheit/Regelungen/IT-Sicherheitsrichtlinie_Cloud-Speicher_2015-11-17.pdf) [accessed: 01.09.2019].
- Universität Bielefeld*, IT-Sicherheitsrichtlinie zur Nutzung von Skype. Version 1.0 (Jun 21, 2012), [http://www.uni-bielefeld.de/informationssicherheit/Regelungen/IT-Sicherheitsrichtlinie\\_Skype\\_2012-06-21.pdf](http://www.uni-bielefeld.de/informationssicherheit/Regelungen/IT-Sicherheitsrichtlinie_Skype_2012-06-21.pdf) [accessed: 01.09.2019].
- Universität Bielefeld*, Blog: uni-intern (8 Jan 8, 2019), [https://ekvv.uni-bielefeld.de/blog/uniintern/entry/geb%C3%Bcndelte\\_expertise\\_f%C3%Bcr\\_effizientes\\_forschungsdaten](https://ekvv.uni-bielefeld.de/blog/uniintern/entry/geb%C3%Bcndelte_expertise_f%C3%Bcr_effizientes_forschungsdaten) [accessed: 01.09.2019].
- Wilkinson, Mark et al.*, The FAIR Guiding Principles for scientific data management and stewardship, in: Scientific Data 3 (2016): 1-9. <https://doi.org/10.1038/sdata.2016.18>.
- Woodward, Kathleen*, The Future of the Humanities – In the Present & in Public, in: Daedalus 138 (2009): 110–123.



# **Navigating Disciplinary Differences in (Digital) Research Projects Through Project Management**

---

*Anna Maria Neubert*

## **Introduction**

In this article I survey our approach to implementing digital research in a collaborative research center by navigating disciplinary differences through tools and methods deriving from project management. I argue that by providing a clear framework and making regulations in cooperative work as well as acknowledging each individual contribution, interdisciplinary collaboration especially in the digital humanities can – even in a short time – produce meaningful and satisfying research results. This article is an account of our strategy to tackle challenges and opportunities that arose during our cooperation in combining the ‘humanities’ and the ‘digital’. It shows where effort paid off and where failures required us to tackle problems and solve them. I conclude by recommending an approach to a new and different recognition of research results and their applicability within disciplinary boundaries that supports a better understanding within interdisciplinary collaborations.



## 1. Implementing digital research in a collaborative research center

Not only has funding of ‘big’ projects<sup>1</sup> in the humanities constantly been growing over the past years,<sup>2</sup> researchers themselves turn to collaboration to explore increasingly complex questions and create large-scale projects. They aim to implement new forms of methodologies that would be either too large or too complex to be completed by a single researcher and therefore are “in need of expertise available from other disciplines.”<sup>3</sup> This increase in quality, depth and scope of research requires a departure from traditional disciplinary boundaries to a more open and fluid concept of humanities research.

Especially in large research collaborations where digital innovations are tested within a humanities environment, the concept of interdisciplinarity is exemplified and tested with each new cooperation on a daily basis. So, “as interdisciplinary collaborations are becoming more common, aligning the interests of computer scientists and humanities scholars requires the formulation of a collaborative infrastructure for research where the approaches, methodologies, pedagogies, and intellectual innovations merge.”<sup>4</sup> Digital humanities (DH) as a field that combines digital concepts with humanities research is an ideal example of challenges, approaches and benefits when disciplinary boundaries are conquered, and collaboration turns out to be a fruitful endeavor for all professions involved. These digital research projects – mostly initiated and “taken on by humanists [...] require manage-

---

1 Cf. for example an article on the emergence of ‘Big Science’: *Weinberg Alvin M.*, Large-Scale Science on the United States, in: *Science*, New Series, Vol. 134, No. 3473 (Jul. 21, 1961), 161–164, <http://www.jstor.org/stable/1708292> [accessed: 01.04.2019].

2 Cf. for only one account from past years: *Allington, Daniel*, The Managerial Humanities; or, Why the Digital Humanities Don’t Exist. (31 Mar. 2013), <http://www.danielallington.net/2013/03/the-managerial-humanities-or-why-the-digital-humanities-dont-exist/> [accessed: 01.04.2019].

3 *Siemens, Lynne*, ‘Faster Alone, Further Together’: Reflections on INKE’s Year Six, in: *Scholarly and Research Communication* 7 (2016), <https://src-online.ca/index.php/src/article/view/250/479> [accessed: 01.04.2019]; *Siemens, Lynne*, ‘More Hands’ means ‘More Ideas’: Collaboration in the Humanities”, in: *Humanities* 4.3 (2015).

4 *Simeone, Michael et al.*, Digging into data using new collaborative infrastructures supporting humanities-based computer science research, in: *First Monday* 16 (2011), <https://firstmonday.org/ojs/index.php/fm/article/view/3372/2950> [accessed: 01.04.2019].

ment”<sup>5</sup>. Thus, “regardless of size, scope, and budget, projects members must coordinate tasks, responsibilities, budgets and achieve objectives.”<sup>6</sup> All those jobs can be tackled and coordinated with tools and methods deriving from project management (PM) that help to issue tasks and responsibilities to eventually achieve satisfying and meaningful (research) results. Unfortunately, the complexity of planning, managing and executing DH projects is still not acknowledged in its entirety and “is usually presented from a beginner’s perspective, offering merely ‘basic principles,’ ‘tips and tricks,’ or ‘top-ten lists’.”<sup>7</sup> However, combining humanistic inquiry with digital approaches allows for a huge variety in questions, implementations and outcomes that all need specific attention and precisely fitting management.

In this article I outline our approach to managing six different digital research projects as part of a pilot phase tasked with evaluating digital methods in various humanities disciplines. The Collaborative Research Center (CRC) 1288 “Practices of Comparing” at Bielefeld University<sup>8</sup> unites 14 humanities research projects – organized within three sections (A, B and C) –, and three central projects that deal with administration (Z), science communication (Ö) and data infrastructure as well as digital humanities (INF). Sub-project INF<sup>9</sup> which coordinated the outlined collaborations in this volume is responsible for providing “data infrastructure and digital humanities” to other projects involved and initiates different forms of digital research over the ongoing first funding period (2017–2020). During the first year of the CRC, INF started a pilot phase to implement digital research methods in existing research projects to augment, extend or renew already existing research questions; thereby questioning how digital research methods can be implemented in existing humanities research and which parameters need altering in order to be able to carry out research that produces valuable results.

---

5 *Ermolaev, Natalia et al.*, Abstract: Project Management for the Digital Humanities, DH2018, Mexico City, <https://dh2018.adho.org/project-management-for-the-digital-humanities/> [accessed: 01.04.2019].

6 *Boyd, Jason/Siemens, Lynne*, Project Management, DHSI@Congress 2014.

7 *Tabak, Edin*, A Hybrid Model for Managing DH Projects, in: Digital Humanities Quarterly 11 (2017), <http://digitalhumanities.org:8081/dhq/vol/11/1/000284/000284.html> [accessed: 01.04.2019].

8 Cf. [http://www.uni-bielefeld.de/\(en\)/sfb1288/index.html](http://www.uni-bielefeld.de/(en)/sfb1288/index.html) [accessed: 01.04.2019].

9 Cf. <http://www.uni-bielefeld.de/sfb1288/projekte/inf.html> [accessed: 01.04.2019].

The pilot phase was planned in advance and provides a number of basic conditions, such as a fixed time frame – one year – and basic defaults and standards about research material that needed digitization. From the beginning it was obvious that good (project) management would be required to deal with mechanics and people<sup>10</sup> throughout the whole process and in order to coordinate requests, challenges and the on-going tasks at hand.

The initial idea that arose at the beginning was that project management can be used as a neutral tool for productive implementation of each of the individual goals. As a basis for collaborative work within and outside of disciplinary boundaries, tools, techniques and methods from project management supported team INF in controlling procedures, keeping the work within deadlines and bringing it to a successful end.

In this article I survey our approach to steer that pilot phase for digital research methods in humanities disciplines ranging from History to Art History to English and German Literature by acknowledging various approaches to research. The contribution at hand is guided by the following questions: How can explicit, mutual expectations in interdisciplinary digital humanities projects be met and managed productively within each disciplinary tradition? Which methods, tools and techniques deriving from project management help to acknowledge individual ideas, pace of work and overall goals? As well as: Can already established benefits from collaborative work such as gaining new skills or new knowledge help pushing cooperation in research projects?

## 2. Definitions: interdisciplinary (research) project(s) in Digital Humanities

As Anthony Paré – Professor Emeritus for Language & Literacy Education at McGill University – points out in a recent blog post, “knowledge-making is a social enterprise that depends on collaborative work.”<sup>11</sup> He thereby captures various topics from a long-lasting yet very current discussion that val-

---

10 *McBride, Melanie*, *Project Management Basics*, New York: Apress, 2016, 2.

11 *Paré, Anthony*, *Scholarship as collaboration: Towards a generous rhetoric.*, <https://doctoralwriting.wordpress.com/2019/02/04/scholarship-as-collaboration-towards-a-generous-rhetoric/#more-2322> [accessed: 01.04.2019].

ues cooperative scholarly work above research as an individual and isolated one-person endeavor and disagrees with the notion of ‘survival of the fittest scholar’ through competition.<sup>12</sup> The collaborative research Paré refers to in his article cannot be carried out without a common understanding of what a group of scholars can achieve within a complex knowledge-making process. But how can a common ground be established, and mutual goals be accomplished in collaborations? And which insights are needed to actually cooperate and pull in the same direction?

Before turning to evaluating our experiences with the implementation of project management methods in a multi-disciplinary context, I am going to elaborate on some fundamental concepts and terminology which we used in planning and executing the pilot phase.

## What defines a project?

Although digital humanities research has already been carried out for many years and is in ever growing demand by funding agencies of all kind, the troubles of carrying out digital research within existing humanities environments continues to be a challenge. The difficulties emerge on levels of mechanics and people; for example, many humanities researchers still see tools, methods and theories originating in the field of DH as “neoliberal and uncritical”<sup>13</sup> assaults. Another critique that often hinders successful collaboration is the attitude towards managerial protocols and the so-called ‘managerial humanities’ that are sometimes used as a synonym for digital humanities research.<sup>14</sup> Nevertheless, it seems obvious, that some kind of management is necessary to do (not only digital, but generally) meaningful research and that fruitful thoughts for research in general often do originate in the planning of research between disciplinary boundaries or by combining disciplines and fields that were not linked beforehand.

In order for those managerial tasks to be carried out successfully, in most cases it helps to think in terms of a ‘project’ – “a temporary endeavour

---

12 Cf. *ibid.*

13 *Svensson, Patrik*, *Big Digital Humanities: Imagining a Meeting Place for the Humanities and the Digital*, Ann Arbor: University of Michigan Press, 2016.

14 Cf. *D. Allington*, *The Managerial Humanities; or, Why the Digital Humanities Don't Exist.*

undertaken to create a unique product, service or result.”<sup>15</sup> Parameters such as time, scope and scale help to grasp a better understanding of the endeavor and support coordinating tasks and responsibilities as well as achieving scheduled and planned objectives. However,

“projects are [also] the way in which human creativity is most effectively harnessed to achieve tangible, lasting results. In the past they may have been called something different, but building a pyramid, painting a ceiling, or funding a nation all required vision, planning and coordinated effort – the essential features of what we now call a project.”<sup>16</sup>

So, when taking this citation seriously, every action done in an academic environment can be seen as a project, even if it is not described as such – from presentations in the first semester to larger tasks such as essays, assignments or dissertations. Research projects, however, require additional attention, as they are comprised of a complex combination of actors and interests. Thus, a research project can be understood as a

- “scientific investigation, usually using scientific methods, to achieve defined objectives.”<sup>17</sup>
- “creative systematic activity undertaken in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this knowledge to devise new applications.”<sup>18</sup>
- “studious inquiry or examination, especially [an] investigation or experimentation aimed at the discovery and interpretation of facts, revision of accepted theories or laws in the light of new facts, or practical application of such new or revised theories or laws.”<sup>19</sup>

---

15 *Project Management Institute*, What is Project Management?, <https://www.pmi.org/about/learn-about-pmi/what-is-project-management> [accessed: 01.04.2019].

16 *Hobbs, Peter*, Project Management (Essential Managers), London: Dorling Kindersley, 2016, 6.

17 *DBpedia*, Project Management, <http://dbpedia.org/ontology/ResearchProject> [accessed: 01.04.2019].

18 *OECD*, Project Management, <https://web.archive.org/web/20070219233912/http://stats.oecd.org/glossary/detail.asp?ID=2312> [accessed: 01.04.2019].

19 *Merriam Webster*, Project Management, <https://www.merriam-webster.com/dictionary/research> [accessed: 01.04.2019].

Also, “projects [in academia] are both nouns and verbs: A project is a kind of scholarship that requires design, management, negotiation, and collaboration. It is also scholarship that projects, in the sense of futurity, as something which is not yet.”<sup>20</sup>

All definitions provided above therefore describe aspects of doing (digital) humanities research. And not only do researchers work in these forms of special activities, they also manage those projects from the very early days on. If a scholar wants to submit a paper, a thesis, or a funding application, she already is coordinating tasks, responsibilities and her working process in order to meet deadline(s). In this light, doing digital research in the form of planned projects is nothing too new, however, the novelty lies in an unprecedented interdependence when collaborating with large numbers of researchers, librarians, research software engineers and other involved stakeholders. Digital Humanities are thus a field that is “most frequently characterized as data- and compute power-intensive, interdisciplinary and highly collaborative in nature.”<sup>21</sup> The underpinning of all projects is nonetheless a framework of specific, but also invariable processes that produce the artifacts and mechanics of project management.

## Varieties of projects in the pilot phase

Subproject INF’s tasks were defined rather broadly in the grant application, and it later became clear that the best way to carry out digital research and introduce digital methods to other involved humanities projects within the CRC, was by forming subprojects that would be manageable and promising with regard to their successful implementation. Thus, after a call for projects, initially six projects<sup>22</sup> from the collaborative research center were selected to take part in the pilot phase for applying and testing digital methods.<sup>23</sup> These

20 *Burdick, Anne et al.*, *Digital\_Humanities*, Cambridge, Mass.:MIT Press, 2012, 124.

21 *Blanke, Tobias/Hedges, Mark/Dunn, Stuart*, Arts and humanities e-science – Current practices and future challenges, in: *Future Generation Computer Systems* 25(2009), 474–480.

22 Involved projects were A04, B01, B03, B05, C01 and C03. More on each project and digital research approach can be found in this whole volume – in much more detail and put in perspective by the humanities researchers themselves.

23 One of the projects was not carried on as the result appeared not to be useful for supporting the research questions already asked.

projects were characterized by a variety and diversity of involved people, contents, material and inputs and their expectations about research goals and outputs. People included PhD students, postdoctoral researchers, Principal Investigators and research assistants who were unified by the fact that they were all humanities researchers, but with backgrounds in History, German Literature, Art History and English Literature. Team INF consisted of computer scientists, librarians, data scientists and digital humanists; consequently, a diverse set of expertise and knowledge was given from the start. However, the projects did not only differ with regard to the people coming from various academic hierarchies, they also differed in terms of contexts within which they were taken on. The individual case studies were either part of PhD projects, initiated questions for research projects within the respective research project in the scope of the CRC, or served as the basis for postdoctoral research projects. In terms of managing these projects, it is important to mention that each project counted as a case study but was always part of a bigger project that was conducted throughout the whole first funding period of the CRC until the end of 2020 of those researchers that tested digital methods in the pilot phase. Goals and intended output were different as well and relied on a variety of source material specific to each project.<sup>24</sup> Nevertheless, as time and (wo)man-power were limited, we tried to offer a technical pipeline that each project could benefit from on different levels.<sup>25</sup>

During the first meetings it soon became obvious that interdisciplinary collaboration is not an easy task, as “it is not clear that all are accustomed to or trained for this type of work.”<sup>26</sup> However, by asking ourselves “Who is involved and brings which qualifications and which knowledge?” and “Which steps are necessary to be successful on the way and in the end?” we were able to recognize and acknowledge disciplinary perceptions for the time being.

---

24 Source materials included French magazines from the sixteenth to the twentieth century, English novels from the seventeenth century or parliamentary debates from twentieth and twenty-first century.

25 For more information on the technical and methodological pipeline that was implemented and the tools that were deployed, see the contribution by Jentsch, Patrick and Stephan Porada “From Text to Data” in this volume.

26 *L. Siemens*, ‘More Hands’ means ‘More Ideas’: Collaboration in the Humanities, 353.

By combining humanities research and the digital, digital humanities collaboration

“requires the creation and deployment of tools for sharing that function to improve collaboration involving large-scale data repository analysis among multiple sites, academic disciplines, and participants through data sharing, software sharing, and knowledge sharing practices.”<sup>27</sup>

These three processes of sharing different parts of those projects helped to form the baseline for managing and coordinating different tasks on the run. However, it was not only obvious that some kind of management was necessary to satisfy all parties involved in this pilot phase, it was also relevant to be open to new or unknown processes that required “skills such as innovation, flexibility, collaboration, communication, negotiation, planning and risk management.”<sup>28</sup> Many of the factors mentioned here were limited by the variety and time of the pilot phase, but it was nevertheless explored which options were available in different situations.

As the definitions and the basic situation were somehow evident and known, they were followed by many processes to shape the phase of management, sharing practices and communication with its different tasks, responsibilities and targeted objectives.

### 3. Project management as the core for navigating disciplinary differences

While researchers and other associated team members welcome collaborations as a way to undertake these kinds of projects, work still needs to be done to prepare individuals for working within a team where interdependent tasks must be coordinated, knowledge and progress must be communicated, and an overall research vision must be accepted and enacted.<sup>29</sup>

---

27 M. Simeone *et al.*, Digging into data using new collaborative infrastructures supporting humanities-based computer science research.

28 L. Siemens, ‘More Hands’ means ‘More Ideas’: Collaboration in the Humanities, 345.

29 L. Siemens, ‘Faster Alone, Further Together’: Reflections on INKE’s Year Six.



These genuinely challenging tasks can be attributed into a so-called recurring cycle of processes that is structured into five groups: (1) Initiating, (2) Planning, (3) Executing, (4) Monitoring and Controlling, and last but not least (5) Closing.<sup>30</sup> While the pilot phase was initiated through a call for projects, the planning phase began by discussing contents, scope and resources with each involved party – mostly consisting of one single researcher but sometimes also of bigger teams of up to four scholars.

However, not only did discussions help in planning various stages of the projects, collected competency and qualifications from various stakeholders helped to draw a larger picture within the planning phase. This expertise for carrying out every aspect of digital research in humanities projects was gained on different levels by team members of INF which prior to working in the collaborative research center were active in various DH projects, university libraries as well as in the field of software development and research management.<sup>31</sup> Furthermore, it was clear that the composition of six different case studies at the same time should provide some basic similarities that align planning and management somewhat better. Consequently, a technological pipeline to digitize texts and carry out digital research methods was implemented for each project to achieve research results that could be utilized by the humanities researcher in their respective larger research projects. As in nearly every DH project, each case study was characterized by being “experimental, modular and incremental”<sup>32</sup> and thus differed from traditional scholarship that was known to the scholars until then.

In order to coordinate the pilot phase and match the tools used for project management, the model shown in figure 1 supported a general understanding of processes that connected and supported the joint effort.

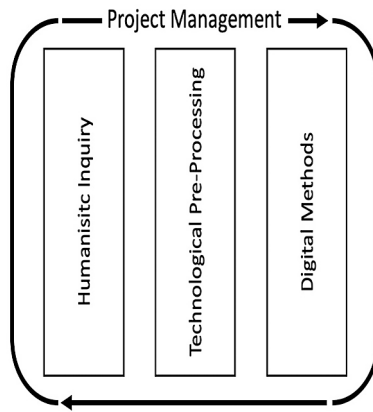
---

30 *Project Management Institute*, What is Project Management?.

31 Cf. for example Neubert (*born Komprecht*), Anna Maria/Röwenstrunk, Daniel, Projektmanagement in digitalen Forschungsprojekten – Ein Leitfaden für interdisziplinäre und kooperative Drittmittelprojekte im Umfeld digitaler Musikedition, in: Kristina Richts/Peter Stadler (eds.), »Ei, dem alten Herrn zoll' ich Achtung gern!«: Festschrift für Joachim Veit zum 60. Geburtstag, München: Allitera, 2016, 509–522.

32 E. Tabak, A Hybrid Model for Managing DH Projects.

Fig. 1: Components of a digital research project linked through project management



The diagram illustrates those three elements on which the research projects were based on. Researchers contributed their research questions on humanities subjects, team INF presented solutions on technological preprocessing and digital methods that could be used to answer the inquiries, and project management aligned all demands to plan and execute every step involved. The model thus can be read from left to right or from right to left, always depending from which angle the projects were seen and taken on. Also, by acknowledging the uniqueness of each individual project and by seeking for similarities in applying steps of technological preprocessing and digital methods it was possible to accomplish a lot of different steps within the set time frame.

For preparation purposes, every project faced the same set of questions which was based on three simple queries which shaped the mechanics of planning different tasks within a project management setting and ended in the basic workflow that was applicable for all projects: “What needs to be done?”, “How is it done?”, and “Who is responsible for which step?”<sup>33</sup> With these questions at hand the planning process was applied in the discussions mentioned beforehand and ended in a diverse mix of methods and tools to implement those research projects.

33 Henderson, Robin, *Research Project Management – Key Concepts* (2010), <https://www.coursehero.com/file/13018002/Key-Concepts-in-Research-Project-Management-Rob-in-Henderson/> [accessed: 31.08.2019], 3.

## Implementing tools and methods for cooperative work

In 2016, Lynne Siemens – one of the leading scholars in doing research on project management in digital humanities settings – writes that processes and principles from project management, like “project plans, reporting structures, knowledge mobilization plans, training, and post-project reporting”<sup>34</sup> are ever more required, “to ensure close alignment between the grant application and the actual outcomes.”<sup>35</sup> And while there already exist some excellent charters that outline basic rules for collaboration in digital humanities projects<sup>36</sup>, they can be helpful for long-lasting projects but were not discussed with the involved researchers in the here described pilot phase due to limited time and a quick start. And although it was not blatantly exposed initially, we applied processes and workflows – actually coming from business-related management – to the case studies from day one; thereby being informed by “information studies and methods in software development, while still being based on values of the humanistic tradition and methods.”<sup>37</sup>

And as Digital Humanities “involves digital objects, digital tools and digital techniques”<sup>38</sup> being brought to bear on traditional humanities scholarship,<sup>39</sup> it was also tried to close this gap by testing other methods and formats coming from science communication and public engagement (see *Inter-*

---

34 Siemens, Lynne, Project management and the digital humanist, in: Constance Crompton/ Richard J. Lane/Ray Siemens (eds.), *Doing Digital Humanities: Practice, Training, Research*, London: Routledge 2016, 343.

35 Cf. *ibid.*

36 Cf. for example the ‘Collaborators’ Bill of Rights’ in ‘Off the Tracks: Laying New Lines for Digital Humanities Scholars’, <http://mcpress.media-commons.org/offthe-tracks/part-one-models-for-collaboration-career-paths-acquiring-institutional-support-and-transformation-in-the-field/a-collaboration/collaborators%E2%80%99bill-of-rights/> [accessed: 01.04.2019] or the ‘Charter’ from the Scholars’ Lab at University of Virginia (Library), <https://scholarslab.lib.virginia.edu/charter/> [accessed: 01.04.2019].

37 E. Tabak, *A Hybrid Model for Managing DH Projects*.

38 Digital objects mainly involved digitized texts in XML/TEI, digital tools were brought in order to support the digitization of those texts and digital techniques were used to test methods like text mining, topic modelling and the building of data bases. More on this can be found in the already mentioned article by Jentsch and Porada.

39 Meeks, Elijah, *The Digital Humanities as Content*, in: Elijah Meeks, *Digital Humanities Specialist* (blog), 19 May 2011, <https://dhs.stanford.edu/the-digital-humanities-as/the-digital-humanities-as-content/> [accessed: 01.04.2019].

*action I: meetings and workshops*). Facing some challenges at the beginning (e. g., a not yet complete team in subproject INF, other projects that needed a final touch before the start of the pilot phase, already existing research projects that needed to be altered to participate as a case study, etc.) we yet always believed that project management could be of use in all situations to encounter common issues that are related to risks, obstacles and tasks<sup>40</sup> which emerge when combining individual ideas and different perceptions of how a research process should proceed. Below follows an outline of the tools and methods that were implemented to meet challenges and seize opportunities to contribute to new questions and research directions in every participating humanities discipline.

### **Assembly of project team(s)**

Before discussing any project related tasks, time tables or methods, each humanities researcher respectively research team was assigned to a collaborative team. Each team consisted of those humanities researchers and selected team members from INF; a computer scientist and a research assistant, a librarian and one of the two digital humanists who managed the collaboration by coordinating stakeholders, tasks and dates. The two Principal Investigators who head team INF – one a digital humanist with a background in medieval history and the other one a computer scientist working in the universities' library – oversaw the collaborations and participated in meetings, workshops and consulted on challenges or occurring failures.

Besides building single research teams for each pilot project, the whole group was understood as one large team to discuss, showcase and present preliminary and possible final results. This cross-team connection was seen as valuable to exchange experience and expert knowledge on humanities subjects but also provided a platform to swap thoughts on using similar (digital) methods and techniques as well as aiming for similar outcomes in different disciplines.

### **Definition of individual research cycle(s)**

Different disciplinary affiliations, as well as one's position within the academic hierarchies may lead to different research cycles over time. While PhD and postdoctoral researchers have a clear time limit set by their tem-

---

40 Cf. L. Siemens, Project management and the digital humanist.

porary employment within the CRC, Principal Investigators and professors are limited through other tasks and research interests within their respective field. That is why it was important to adjust the collaborative research plans as much as possible to secure the benefits of working together and built strong working relationships despite of time constraints or boundaries that are applied by disciplinary, departmental, faculty or other systems in which every researcher is integrated throughout the academic environment.

Furthermore, it was anticipated that the research cycle of projects taken on in the pilot phase must be seen within the bigger picture of the four-year funding phase for the CRC in general. It was tried to acknowledge the presence of individual situations and the facilitation of a different pace of work, however, meetings in the larger group and discussions on intermediate results were always synced in order to disclose common misunderstandings or align changes in the project plan.

### Project plans

Although it was not quite as easy in the beginning, we drafted two project plans. First, a project plan was designed for the whole endeavor from and for team INF solely, to coordinate and align tasks that were part of each individual team member and second, project plans for each individual project were created, to adjust scope, scale and involvement which differed if (only) slightly from project to project. In general, we adopted a so-called rolling wave plan<sup>41</sup> that allowed filling in aspects of the projects on “a rolling basis”<sup>42</sup> and thus made it possible to adjust to occurring challenges and risks.

The plans showed the general allocation of time within the set time frame of a year and represented upcoming tasks in the style of *Kanban*<sup>43</sup> workflows. Each planning step was available through the online project management system *Redmine*<sup>44</sup> (see *Documentation*) and roughly followed the concept of work breakdown structure (WBS)<sup>45</sup> which allowed a detailed planning of every step.

---

41 R. Henderson, *Research Project Management – Key Concepts*, 4.

42 Cf. *ibid.*

43 Atlassian, *Kanban*, <https://www.atlassian.com/agile/kanban> [accessed: 01.04.2019].

44 Cf. <https://www.redmine.org/> [accessed: 10.02.2019].

45 Cf. R. Henderson *Research Project Management – Key Concepts*, 4.

Each project plan can only be successful with accompanying steps that allow for a better understanding and alignment of the common project goal. And while a project plan is one of the basics for the execution of a project, we drew on more methods. Norms on how to behave in meetings, when working together or communicating<sup>46</sup> were also considered as were new ideas from user design or science communication.

## Visualizations

Especially when talking about technical solutions for or realizations of humanistic inquiry with humanities scholars who – ordinarily – are used to text as their main research material, the idea of drawing, wire framing and plotting out steps of technological processes comes in handy. In our pilot phase, the process of “Thinking through Practice”<sup>47</sup> helped to start a discussion about how to structure and model material and data in order to achieve a satisfying result. This does not always need to be in a digital format, drawing on white boards or flip charts together supported the communication and the finding of a consensus on how to tackle challenges and proceed with each project. There are many forms of visual methods deriving from design that add another layer of knowledge and prove to lead to a synchronization of how to move forward; thus, we used techniques like mind mapping or brain storming as forms of visualizing research processes. This was not only realized in smaller team meetings but also in forms of workshops.

## Interaction I: meetings and workshops

Meetings took place in smaller research teams as well as in the large group. While there was always a set of standard questions prepared for those regular meetings in each project team, workshops in the large team throughout had a topic that helped to discuss each stage of the individual projects. As we tried to synchronize technological and methodological steps in order to provide opportunities for everyone to merge research progress or discuss challenges, it was quite fruitful to prepare conversation guidelines for the smaller meetings but also include new formats of collaborations in the larger gatherings. Innovative ideas from science communication, agile software

---

46 L. Siemens, *Project management and the digital humanist.*, 352.

47 Duxbury, Lesley/Grierson, Elizabeth M./Waite, Dianne (eds.), *Thinking Through Practice: Art as Research in the Academy*, Melbourne: RMIT Publishing, 2007.

development and information studies proved to be quite useful for getting researchers to talk and meet different standards in a very limited time. World cafés,<sup>48</sup> booksprints<sup>49</sup> or expert discussions were those formats which had a huge impact on how the collaboration moved forward. They helped to learn from each other, respect time constraints of each individual and meet up to the standards that were set from the beginning.

### Interaction II: human-machine-interaction

But not only did meetups with real human beings help us understand technical and methodological challenges, workshops on how to interact with the machine were valuable to disseminate a better understanding of how the computer works and is used for the particular research questions posed by the humanities researcher. In those meetings and workshops, it was generally introduced how the technical preprocessing pipeline was employed and worked (or precisely did not work) for each project and how researchers themselves could learn to handle digital methods. With applications like *AntConc*,<sup>50</sup> *Voyant*<sup>51</sup> or *Mallet*<sup>52</sup> we introduced tools that can be used in order to enable researchers to work with their material alone and beyond the pilot phase. One goal of the pilot phase was always to work beyond the team effort and facilitate new skills when employing digital innovations and make those innovations usable for research processes in the humanities.

### Documentation

Using the flexible project management web application *Redmine*<sup>53</sup> helped to coordinate each project and provided a platform for communicating with the researchers as well as documenting important steps along the way. Each project involved in the pilot phase was assigned an own subproject that always had a similar structure. Each meeting, discussion, and decision was

---

48 *The World Café*, <http://www.theworldcafe.com/key-concepts-resources/world-cafe-method/> [accessed: 01.04.2019].

49 Cf. for example *Zennaro, Marco et al.*, *Book Sprint: A New Model for Rapid Book Authoring and Content Development*, in: *International Journal of the Book* 4 (2007), 105–109. <http://ijb.cgpublisher.com/product/pub.27/prod.120> [accessed: 01.04.2019].

50 Cf. *Antcon*, <http://www.laurenceanthony.net/software/antcon/> [accessed: 01.04.2019].

51 Cf. *Voyant* – see through texts, <https://voyant-tools.org/> [accessed: 01.04.2019].

52 Cf. *Mallet* – Topic Modeling, <http://mallet.cs.umass.edu/topics.php> [accessed: 01.04.2019].

53 Cf. *Redmine*, <https://www.redmine.org/> [accessed: 01.04.2019].

documented in the wiki in order to be consulted later on, and important documents were stored there as well.<sup>54</sup> While we did not make use of the whole project management functions Redmine offers, such as a ticketing system, it served as a communication platform for each step. It supported the execution and secured monitoring for each team member involved. Documentation already proves to be quite helpful as of now – two years later – as all stakeholders are able to draw on those materials deposited in the wiki which supports in, for example, writing the articles about the whole pilot phase.

The synopsis of tools and methods depicted above can only be a selection of what we drew upon throughout but represents the most important facets of our project management. While working in interdisciplinary collaborations is a constant challenge, it can be met with preliminary work to seize the chance of developing something novel to each community. It is absolutely clear that some kind of planning is necessary to implement cooperative research and the methodological mix we chose proved to be quite fruitful.

### **Co-creation: credit where credit is due**

One issue that is often encountered when working in teams is the credibility for different steps along the way. As humanities research still has a “historical emphasis on the single author”,<sup>55</sup> a team-based approach to research challenges traditional humanities. Individual authorship<sup>56</sup> however becomes quite a challenge as one person cannot possibly carry out all tasks which would be necessary to come to a satisfying and meaningful result in the end. The differentiation of content consumption, content creation, and content management needs to be applied to the roles of collaborative teams in digital humanities projects as soon as it becomes explicit who takes responsibility for which task. So, “the most effective digital humanities work is done when a scholar has an innovative, sophisticated agenda that can be furthered by application of computational methods [and] digital publication.”<sup>57</sup> This

---

54 Research material was exchanged via the university-cloud ‘sciebo’, as we wanted to split the organization and communication platform from the working environment.

55 L. Siemens, Project management and the digital humanist, 352.

56 A. Burdick et al., Digital\_Humanities, 125.

57 Meeks, Elijah, How Collaboration Works and How It Can Fail, in: Elijah Meeks, Digital Humanities Specialist – humanities software, visualization and analysis (blog), 27 May 2013, <https://dhs.stanford.edu/natural-law/how-collaboration-works-and-how-it-can-fail/> [accessed: 01.04.2019].



means that each involved party should get credit for the contribution to each part of the project. While there is no standard of new crediting systems yet, there is a “trend toward the differentiation of roles such as principal investigator, researcher, designer, programmer, modeler, editor, and the like.”<sup>58</sup> It should be discussed which roles are involved in which process and how even ‘invisible’ jobs – like project management in most of the cases – can get credit for the fulfilled jobs. We opted for an approach that tries to combine the individual authorship and the team-based approach. In producing the collection of the articles at hand, each humanities researcher (team) is able to contribute a paper that is authored individually, however, on the website the data publications, data stories and research results pay tribute to the whole researcher and developer team involved.

#### 4. Challenges and failures – a way to succeed in the end?

Any project – and maybe research projects in particular – requires careful preliminary planning to come into being. Yet, no amount of planning can prevent unforeseen developments or guarantee a smooth ride, from beginning to end. It is thus known that “there will be bumps in the road.”<sup>59</sup> Common failure characteristics include, for example, slipped schedules, significant amounts of firefighting, which means that much time is spent on unanticipated problems, final results turn out to not come close to the original expectations, surprising decisions by any team member, failure to meet compliance requirements or late realization that the team cannot deliver on time.<sup>60</sup>

Needless to say, we faced some of those bumps along our own way. Starting out, we wanted to make use of existing and well established tools that had already been implemented by the digital humanities community.<sup>61</sup> Unfortunately, as it turns out, most of the tools that were funded in the past lack continued and continuous (technical) support after their initial funding periods

---

58 A. Burdick *et al.*, *Digital Humanities*, 125.

59 M. McBride *Project Management Basics*, 118.

60 Cf. M. McBride, *Project Management Basics*, 118.

61 TextGrid and Weblicht for example are some of those tools that we tried to introduce but failed to do so in the end.

and were never adjusted to current technical developments. This means, we had the option to either enhance those tools by ourselves or find another way to run different applications consecutively. This process required a high investment of time as testing and adaptation of tools to our projects' needs and contexts turned out to be a challenging activity. Furthermore, most of the tools we wanted to use were somewhat 'techie' and thus an obstacle for most humanities' researchers involved. As all are used to functioning (mostly proprietary) software on very high levels it does not come easy to scale down and chum up with applications that need to be addressed and used differently. We tried to scale down the usage of such tools as much as possible and provided scholars with tools that could be handled easier.

Another obstacle that came up during the first few months was that either materials were too hard to process (e. g., no OCR was possible due to bad scans) or other materials were discovered which were suited better for the pilot phase. We therefore scaled down on different projects<sup>62</sup> and changed the focus for some other ones at that time. Moreover, after applying several steps of preprocessing, it was apparent that one project could definitely not use those results produced by digital methods. After considering and testing other methods and discussing other options it became obvious that perceptions and ideas for the pilot phase differed, and together as a team we decided to terminate this one project.

Another challenge we faced as a team did not primarily concern the working environment but social interactions within the team. As team members were trained in different cultures and had qualifications in various areas, it was not always easy to understand each other on a content level, and it sometimes felt – like in many other digital humanities projects – “that oftentimes collaboration with computer scientists [is] more like colonization by computer scientists.”<sup>63</sup> We soon realized that working processes, communication practices and general disciplinary standards vary widely and conflicts must be addressed as soon as they appear. It was not clear at first how much time it would cost to translate these differing perceptions

---

62 Cf. for example the article by Heyder, Joris C. “Challenging the *Copia*. Ways to a Successful Big Data Analysis of Eighteenth-Century Magazines and Treatises on Art Connoisseurship” in this volume.

63 Meeks, Elijah, Digital Humanities as a Thunderdome, in: Journal of Digital Humanities 1, <http://journalofdigitalhumanities.org/1-1/digital-humanities-as-thunderdome-by-eljah-meeks/> [accessed: 01.04.2019].

and how frequent discussions would revolve around ways of getting everyone on the same page. Nevertheless, we always tried to talk things out and not repeat errors made by other projects, like Project Bamboo the “greatest impediment [of which] was the lack of a shared vision among project leaders, development teams, and communications staff.”<sup>64</sup> In the end, we were always able to gather around the same perception and work on the collective goals together. However, by only just describing these issues it becomes evident that collaboration needs an endless willingness to cooperate and work on arising challenges together.

## 5. Lessons learned – creating unique digital research projects as temporary endeavors

The preceding chapters have demonstrated that there is a lot of potential in doing planned digital research within limited boundaries such as collaborative research centers many humanities disciplines are already associated with from the start. By re-using already existing tools and doing digital humanities research with a “low end DH”<sup>65</sup> – or “minimalist understanding”<sup>66</sup> approach, a variety of outcomes can be realized in a short time. Interdisciplinary cooperation is usually a complex and challenging endeavor, but the effort pays off and opens up new spaces for research in between disciplinary boundaries and furthers innovative and thought-provoking ideas.

However, there are some lessons we learned during the execution of this pilot phase which need to be addressed during a next funding phase or any other project that is taken on by a diverse team composed of various professions and characters.

---

64 *Dombrowski, Quinn*, What Ever Happened to Project Bamboo?, in: *Literary and Linguistic Computing* (2014). Published by Oxford University Press on behalf of EADH. doi:10.1093/lc/fqu026.

65 *Burghardt, Manuel/Wolff, Christian*, Digital Humanities: Buzzword oder Strukturwandel der Geisteswissenschaften?, in: *Blick in die Wissenschaft – Forschungsmagazin der Universität Regensburg* 29 (2014), S. 40, [https://dhregensburg.files.wordpress.com/.../wolff-burghardt-seiten-aus-biw\\_46-5.pdf](https://dhregensburg.files.wordpress.com/.../wolff-burghardt-seiten-aus-biw_46-5.pdf) [accessed: 01.04.2019].

66 *Kirsch, Adam*, Technology Is Taking Over English Departments – The false promise of the digital humanities, in: *New Republic* 2 (2014), <https://newrepublic.com/article/117428/limits-digital-humanities-adam-kirsch> [accessed: 01.04.2019].

- (1) It is never too early to consider project management methods for implementing digital humanities projects. Not only does an initiating and planning phase prevent hasty research, it also gives every involved stakeholder some time to review aspects and parameters of the designated research team, research cycle and desired outcome. Furthermore, there is a chance that so-called “bumps in the road” will be discovered beforehand and resolved in time to still achieve satisfying results.
- (2) Acknowledgment of different cultures of the involved disciplines is an important issue that needs to be addressed from the start. Discussing the various training and qualification standards in the arts and humanities, the social sciences and the sciences can help to nurture a mutual understanding that supports fruitful collaborations. The only chance to prevent misconceptions and the undue preference of any party is by clarifying the (learned) conceptions and keeping an on-going conversation alive to discuss similarities and differences so as to find a common path.
- (3) Scholarly communication has to be expanded to not only discuss research within the known and learned disciplinary boundaries, but also with other researchers from other disciplines, fields or cultures and also with a wider public.<sup>67</sup> Though interdisciplinary cooperation can support this way of communicating it should not be ignored that an ongoing conversation is not always as easy to keep up as it may seem. One chance arising from this is to discuss conceptions and expectations through produced (intermediate) research results and new research questions that can combine disciplinary claims and standards.
- (4) While it may be difficult to completely ignore academic hierarchies, it is certainly advisable to try to set them aside and thereby create a cooperation that recognizes skills and knowledge independently from age or standing within the academic community or in a university setting. Especially, when applying digital innovations, younger team members may have a different approach to dealing with challenges and can be motivated to think outside the box. Paired with the knowledge of scholars who have been in the system for a while, the results can exceed what

---

67 Also, through publication practices, see the article by Schlicht, Helene “Open Access, Open Data, Open Software? Proprietary Tools and Their Restrictions” in this volume.

would be possible individually. Also, by integrating developers, librarians and archivists, an intersection through the whole university can be created to benefit from every profession's specific knowledge and experience.

- (5) The acceptance of varying paces of work as well as the acknowledgement that visibility of intermediate stages during the research cycle can differ among persons, disciplines and over time. Not every person works the same way and particularly in an academic setting it is not always obvious how one step follows the other. As long as there is a project plan with agreed deadlines, it should be acknowledged that people do research in very different ways but produce satisfying results, nevertheless.

All in all, each challenge that occurred during the collaboration helped to find useful ways of dealing with each problem. One aspect that we expect to be very valuable for further projects is to reflect on each collaboration. Through semi-structured interviews and additional discussions with all researchers involved we hope to optimize processes and outcomes to fit digital research cycles into a larger research context. Not only do already established sets of questions help to learn from other projects and prepare a better understanding for all parties from the start (see figure 2), it is also worth striving for norms and making each team member accountable from the start.<sup>68</sup>

By drawing on proposals from the well-established community of digital humanities practitioners, participants can be assured that generally accepted benefits from working in collaborative research teams can – up to a certain point – be achieved. Moreover, through the encounter of new technologies and as yet unknown research novel career paths can be pursued to enrich traditional humanities research.<sup>69</sup> As such it seems that the benefits of gaining new skills and creating new knowledge through cooperation in research settings cannot be stressed enough and that to plan and execute each step on the way secures improvement for researchers themselves, whole fields and disciplines and thus research itself.

---

<sup>68</sup> L. Siemens, Project management and the digital humanist, 351.

<sup>69</sup> Cf. for example “alt-academy – Alternative Academic Careers for Humanities Scholars”, available at <http://mediacommons.org/alt-ac/about> [accessed: 01.2019].

*Fig. 2: Unsworth's so-called deformation of questions from the Human Genome project to discover common experimental methods and problems in digital humanities projects which help to plan a project from the start.<sup>70</sup>*

1. Queries: What questions will you want to answer? What types of data will you need to answer these questions? Which of these data types are permanent, which are temporary but important, and which will need to be regularly updated? What uses will you have for generic data in the next 5 years?
2. Tools: What protocols and tools for data submission, viewing, analysis, annotation, curation, comparison, and manipulation will you need to make maximal use of the data? What sorts of links among datasets will be useful?
3. Infrastructure: What critical infrastructures will be needed to support the queries you want to perform and what attributes should these infrastructures have? In what ways should they be flexible, and how should they stay current? How should they be maintained?
4. Standards: What kind of community-agreed standards are needed, e.g. controlled vocabularies, datatypes, annotations, and structures? How should these be defined and established?

## 6. Conclusion: What actually counts as a result?

One important lesson we learned during the pilot phase was about the differing conceptions of what a research result in different disciplines looks like. Confronted with questions surrounding what a meaningful and/or satisfying research result is (One that answers the question posed, or one that challenges views and expectations? Or only one that sparks new questions and directs further research?), it was obvious, that even if interdisciplinary collaboration appears to be successful for the stakeholders involved, it also depends on the interpretation of each party.

As different trainings, qualifications and disciplinary standards exist, the meaning of results was discussed with scholars themselves – especially in those cases where digital methods were applied. Together, we came up with the statement that every result produced through technological processing of the researched materials is a result that can be interpreted and used as a fruitful representation of the larger research done in each project. Even if an outcome cannot be interpreted within disciplinary boundaries, it alerts the researcher to a different picture and challenges perceptions and

---

70 Cf. *Unsworth, John*, Scholarly Primitives, <http://www.people.virginia.edu/~jmuzm/Kings.5-00/primitives.html> [accessed: 01.04.2019].

apprehensions.<sup>71</sup> So, even though Digital Humanities' "core commitments [are to] harmonize with the long-standing values of the humanistic tradition: the pursuit of analytical acuity and clarity, the making of effective arguments, the rigorous use of evidence, and communicative expressivity and efficacy",<sup>72</sup> it melds with "hands-on work with vastly expanded data sets, across media and through new couplings of the digital and the physical, resulting in definitions of and engagements with knowledge that encompass the entire human sensorium."<sup>73</sup> Thus, one of the biggest lessons learned is that it is valuable and promising to already include imaginings of results and their applicability within each context in the initiating and planning phase. Mutual expectations of each involved researcher, developer and project manager will consequently gain another angle to be implemented productively within each disciplinary tradition.

One of the researchers interviewed by Lynne Siemens on interdisciplinary collaboration points out that "while it is faster to do things alone, it is possible to go further when working in a team."<sup>74</sup> This quote comes very close to our experience during the pilot phase. By applying methods and tools from various fields to organize a testing environment for digital research in the humanities we were successful in acknowledging individual ideas, pace of work and overall goals in most of the cases. The contributions collected in this volume highlight different approaches to examine digital methods in each humanities discipline and demonstrate the wide variability of how results can and also should be interpreted when combining 'humanistic inquiry' and the 'digital'.

---

71 Cf. for example the article by Peters, Christine "Text Mining, Travel Writing, and the Semantics of the Global: An AntConc Analysis of Alexander von Humboldt's *Reise in die Äquinoktial-Gegenden des Neuen Kontinents*" in this volume where long-lasting research perceptions are challenged by the result of her interpretation.

72 A. *Burdick et al.*, *Digital\_Humanities*, 124.

73 Cf. *ibid.*, 124.

74 Siemens, Lynne. "'Faster Alone, Further Together': Reflections on INKE's Year Six", in *Scholarly and Research Communication*. Vol 7 No 2/3 (2016), <https://src-online.ca/index.php/src/article/view/250/479> [accessed: 01.04.2019].

## Bibliography

- Allington, Daniel*, The Managerial Humanities; or, Why the Digital Humanities Don't Exist. (31 Mar. 2013), <http://www.danielallington.net/2013/03/the-managerial-humanities-or-why-the-digital-humanities-dont-exist/> [accessed: 01.04.2019].
- Atlassian*, Kanban, <https://www.atlassian.com/agile/kanban> [accessed: 01.04.2019].
- Blanke, Tobias/Hedges, Mark/Dunn, Stuart*, Arts and humanities e-science – Current practices and future challenges, in: *Future Generation Computer Systems* 25(2009)
- Boyd, Jason/Siemens, Lynne*, Project Management, DHSI@Congress 2014.
- Burdick, Anne et al.*, *Digital Humanities*, Cambridge, Mass.: MIT Press, 2012.
- Burghardt, Manuel/Wolff, Christian*, Digital Humanities: Buzzword oder Strukturwandel der Geisteswissenschaften?, in: *Blick in die Wissenschaft – Forschungsmagazin der Universität Regensburg* 29 (2014), S. 40, [https://dhregensburg.files.wordpress.com/.../wolff-burghardt-seiten-aus-biw\\_46-5.pdf](https://dhregensburg.files.wordpress.com/.../wolff-burghardt-seiten-aus-biw_46-5.pdf) [accessed: 01.04.2019].
- 'Charter' from the Scholars' Lab at University of Virginia (Library), <https://scholarslab.lib.virginia.edu/charter/> [accessed: 01.04.2019].
- 'Collaborators' Bill of Rights' in 'Off the Tracks: Laying New Lines for Digital Humanities Scholars', <http://mcpres.media-commons.org/offthe-tracks/part-one-models-for-collaboration-career-paths-acquiring-institutional-support-and-transformation-in-the-field/a-collaboration/collaborators%E2%80%99-bill-of-rights/> [accessed: 01.04.2019].
- DBpedia*, Project Management, <http://dbpedia.org/ontology/ResearchProject> [accessed: 01.04.2019].
- Dombrowski, Quinn*, What Ever Happened to Project Bamboo?, in: *Literary and Linguistic Computing* (2014). Published by Oxford University Press on behalf of EADH. doi:10.1093/llc/fqu026.
- Duxbury, Lesley/Grierson, Elizabeth M./Waite, Dianne* (eds.), *Thinking Through Practice: Art as Research in the Academy*, Melbourne: RMIT Publishing, 2007.
- Ermolaev, Natalia et al.*, Abstract: Project Management for the Digital Humanities, DH2018, Mexico City, <https://dh2018.adho.org/project-management-for-the-digital-humanities/> [accessed: 01.04.2019].



- Henderson, Robin*, Research Project Management – Key Concepts (2010), <https://www.coursehero.com/file/13018002/Key-Concepts-in-Research-Project-Management-Robin-Henderson/> [accessed: 31.08.2019].
- Hobbs, Peter*, Project Management (Essential Managers), London: Dorling Kindersley, 2009.
- Kirsch, Adam*, Technology Is Taking Over English Departments – The false promise of the digital humanities, in: New Republic 2 (2014), <https://newrepublic.com/article/117428/limits-digital-humanities-adam-kirsch> [accessed: 01.04.2019].
- McBride, Melanie*, Project Management Basics, New York: Apress, 2016.
- Meeks, Elijah*, Digital Humanities as a Thunderdome, in: Journal of Digital Humanities 1, <http://journalofdigitalhumanities.org/1-1/digital-humanities-as-thunderdome-by-elijah-meeks/> [accessed: 01.04.2019].
- Meeks, Elijah*, How Collaboration Works and How It Can Fail, in: Elijah Meeks, Digital Humanities Specialist – humanities software, visualization and analysis (blog), 27 May 2013, <https://dhs.stanford.edu/natural-law/how-collaboration-works-and-how-it-can-fail/> [accessed: 01.04.2019].
- Meeks, Elijah*, The Digital Humanities as Content, in: Elijah Meeks, Digital Humanities Specialist (blog), 19 May 2011, <https://dhs.stanford.edu/the-digital-humanities-as/the-digital-humanities-as-content/> [accessed: 01.04.2019].
- Merriam Webster*, Project Management, <https://www.merriam-webster.com/dictionary/research> [accessed: 01.04.2019].
- Neubert (born Komprecht), Anna Maria/Röwenstrunk, Daniel*, Projektmanagement in digitalen Forschungsprojekten – Ein Leitfaden für interdisziplinäre und kooperative Drittmittelprojekte im Umfeld digitaler Musikedition, in: Kristina Richts/Peter Stadler (eds.), »Ei, dem alten Herrn zoll' ich Achtung gern!«: Festschrift für Joachim Veit zum 60. Geburtstag, München: Allitera, 2016, 509–522.
- OECD*, Project Management, <https://web.archive.org/web/20070219233912/http://stats.oecd.org/glossary/detail.asp?ID=2312> [accessed: 01.04.2019].
- Paré, Anthony*, Scholarship as collaboration: Towards a generous rhetoric., <https://doctoralwriting.wordpress.com/2019/02/04/scholarship-as-collaboration-towards-a-generous-rhetoric/#more-2322> [accessed: 01.04.2019].
- Project Management Institute*, What is Project Management?, <https://www.pmi.org/about/learn-about-pmi/what-is-project-management> [accessed: 01.04.2019].

- Siemens, Lynne*, 'Faster Alone, Further Together': Reflections on INKE's Year Six, in: *Scholarly and Research Communication* 7 (2016), <https://src-online.ca/index.php/src/article/view/250/479> [accessed: 01.04.2019]
- Siemens, Lynne*, Project management and the digital humanist, in: Constance Crompton/Richard J. Lane/Ray Siemens (eds.), *Doing Digital Humanities: Practice, Training, Research*, London: Routledge 2016, 343.
- Siemens, Lynne*, 'More Hands' means 'More Ideas': Collaboration in the Humanities", in: *Humanities* 4.3 (2015).
- Simeone, Michael et al.*, Digging into data using new collaborative infrastructures supporting humanities-based computer science research, in: *First Monday* 16 (2011), <https://firstmonday.org/ojs/index.php/fm/article/view/3372/2950> [accessed: 01.04.2019].
- Svensson, Patrik*, *Big Digital Humanities: Imagining a Meeting Place for the Humanities and the Digital*, Ann Arbor: University of Michigan Press, 2016.
- Tabak, Edin*, A Hybrid Model for Managing DH Projects, in: *Digital Humanities Quarterly* 11 (2017), <http://digitalhumanities.org:8081/dhq/vol/11/1/000284/000284.html> [accessed: 01.04.2019].
- The World Café*, <http://www.theworldcafe.com/key-concepts-resources/world-cafe-method/> [accessed: 01.04.2019].
- Unsworth, John*, Scholarly Primitives, <http://www.people.virginia.edu/~jmu2m/Kings.5-00/primitives.html> [accessed: 01.04.2019].
- Weinberg, Alvin M.*, Large-Scale Science on the United States, in: *Science, New Series*, Vol. 134, No. 3473 (Jul. 21, 1961), 161–164, <http://www.jstor.org/stable/1708292> [accessed: 01.04.2019].
- Zennaro, Marco et al.*, Book Sprint: A New Model for Rapid Book Authoring and Content Development, in: *International Journal of the Book* 4 (2007), 105–109.



## **II. From Text to Data**



# From Text to Data

## Digitization, Text Analysis and Corpus Linguistics

---

*Patrick Jentsch, Stephan Porada*

### 1. Introduction

Working with sources like books, protocols and other documents is the basis of most scientific work and research in the humanities. Unfortunately, most of these sources are only available on paper or come in other analog forms like parchments.

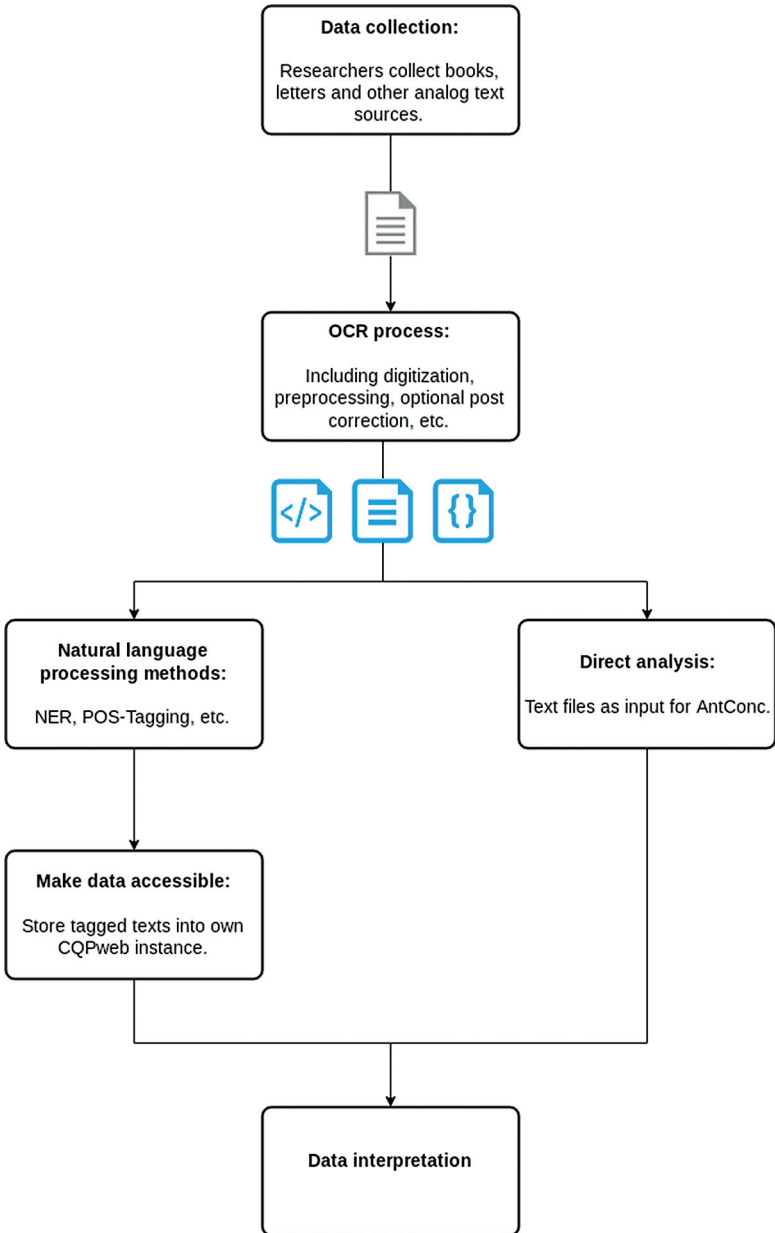
Within the field of Digital Humanities methods like data mining, text analysis and corpus linguistics are widely spread and used. To apply these methods to current historical research, historic text sources have to be digitized and turned into machine readable data by following the main steps outlined below.

At first, sources must be scanned to create digital representations of them. Following this, their images are used as input for our optical character recognition (OCR) pipeline, which produces plain text data. This text can then be further analyzed by means of the methods mentioned above. The analysis involves several natural language processing methods (NLP) which will also be discussed. Figure 1 shows the overall process from text to data, namely the main steps of data collection, OCR processing, data analysis and data interpretation.

The goal of this article is to explain the technologies and software used by the INF team (Data Infrastructure and Digital Humanities) of the Collaborative Research Center (SFB) 1288 “Practices of Comparing” to turn historical documents into digitized text and thus create the data basis for further research steps including text analysis and corpus linguistics.

In part two we present arguments advocating the use of free and open source software (FOSS). Part three is an overview of the basic software and

Fig. 1: Flow chart showing the entire process from text to data



technologies used to implement and to deploy our pipelines to production. The fourth part is a detailed description of the OCR pipeline, its software and internal processes. The last part discusses the different natural language processing (NLP) and computer linguistic methods which can be applied to the output texts of the OCR pipeline. Most of these methods are implemented by using *spaCy*.

The source code of the pipelines described in part four and five can be downloaded from the Bielefeld University's *GitLab* page.<sup>1</sup>

Both repositories include detailed instructions for the installation and usage of both pipelines.

## 2. Free and open-source software (FOSS)

One of the main goals besides turning text into data is to only use software that meets specific criteria in terms of sustainability, longevity and openness. These selection criteria are based on the “Software Evaluation Criteria-based Assessment”<sup>2</sup> guideline published by the Software Sustainability Institute.<sup>3</sup> The latter helped us decide on which software suited our needs best. The following paragraphs show and explain some of our main selection criteria.

---

1 The repository [https://gitlab.uni-bielefeld.de/sfb1288inf/ocr/tree/from\\_text\\_to\\_data](https://gitlab.uni-bielefeld.de/sfb1288inf/ocr/tree/from_text_to_data) contains the OCR pipeline. The repository [https://gitlab.uni-bielefeld.de/sfb1288inf/nlp/tree/from\\_text\\_to\\_data](https://gitlab.uni-bielefeld.de/sfb1288inf/nlp/tree/from_text_to_data) contains the NLP pipeline used for POS (part-of-speech) tagging, NER (named entity recognition) tagging, etc. [accessed: 31.08.2019].

2 Jackson, Mike/Crouch, Steve/Baxter, Rob, Software Evaluation: Criteria-Based Assessment (Software Sustainability Institute, November 2011), <https://software.ac.uk/sites/default/files/SSI-SoftwareEvaluationCriteria.pdf> [accessed: 31.08.2019].

3 This guideline again is based on the *ISO/IEC 9126-1* standard. The standard has been revised and was replaced in 2011 by the new *ISO/IEC 25010:2011* standard. *International Organization for Standardization*, ISO/IEC 9126-1:2001: Software Engineering – Product Quality – Part 1: Quality Model (International Organization for Standardization, June 2001), <https://www.iso.org/standard/22749.html> [accessed: 31.08.2019] and *International Organization for Standardization*, ISO/IEC 25010:2011: Systems and Software Engineering – Systems and Software Quality Requirements and Evaluation (SQuaRE) – System and Software Quality Models (International Organization for Standardization, March 2011), <https://www.iso.org/standard/35733.html> [accessed: 31.08.2019].



Sustainability, maintainability and usability ensure that every step in the process of turning text into data is documented and therefore reproducible. The main selection criteria consist of different subcriteria. To assess the sustainability and maintainability of software, for example, questions regarding copyright and licensing have to be answered.<sup>4</sup> Choosing free and open-source software (FOSS) ensures that the source code of those tools can always be traced. Processing steps conducted with FOSS are therefore always documented and reproducible. Another example of a subcriterion is interoperability.<sup>5</sup> The main aim of this criterion is to ensure that the software is easily interoperable with other software. In our case we mainly wanted to ensure that every software produces data output in open standards like XML or just plain text. This is crucial because output data created by one application or one software has to be easily usable with other software. In addition, open formats like XML are user-friendly and can easily be read for first evaluations of the data. Therefore, it is best practice to publish data in formats like XML because the scientific community can easily review, use and alter the data.

Conducting a criteria based software assessment using the main and subcriteria mentioned above naturally leads to only or mainly choosing and using FOSS.

In addition to the already mentioned advantages, this process ensures that only software is chosen that can be used over longer periods of time, even beyond the actual project phase. This is necessary because software and services which are based on a particular program still have to be usable after the end of the project phase.

### 3. Basic software

Our software implementations are based on some basic technologies. This part gives a brief introduction to this software.

---

4 M. Jackson/S. Crouch/R. Baxter, *Software Evaluation*, 7-8.

5 *Ibid.*, 13.

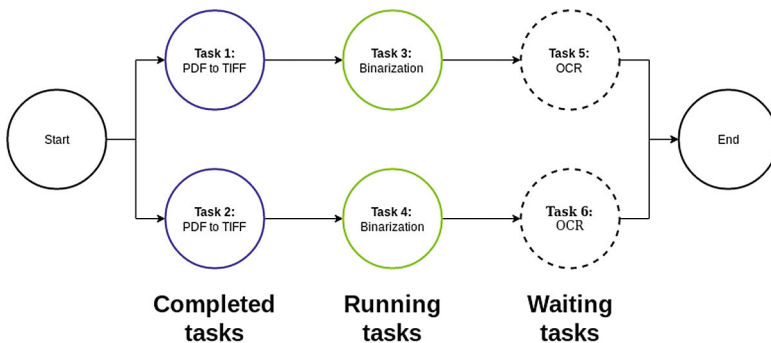
## pyFlow

The OCR and NLP modules we developed were implemented by using the python module *pyFlow*.<sup>6</sup> It is currently only available for Python 2.<sup>7</sup>

This module is used to manage multiple tasks in the context of a task dependency graph. In general this means that *pyFlow* creates several tasks with hierarchical dependencies (upstream tasks have to be completed before the dependent task will be executed). Tasks with the same satisfied dependency (or tasks with no first dependencies) can be completed concurrently. Other tasks depending on those to be finished first will only be executed if the necessary dependencies are satisfied. Figure 2 shows an example of a simple OCR process.

At first, the OCR pipeline has to convert input files from PDF to TIFF because the binarization, which constitutes the following preprocessing step, can only be done with TIFF files. The conversion of each PDF file to TIFF is a single task. Because *pyFlow* is designed with parallelization in mind, some tasks can be executed concurrently. In this case, the conversion from PDF to TIFF can be done for multiple files simultaneously depending on the available RAM and CPU cores.<sup>8</sup> As soon as all PDFs are converted, the binarization of the files can start.

Fig. 2: Dependency graph example for a simple OCR process



6 Saunders, Chris, *PyFlow: A Lightweight Parallel Task Engine*, version 1.1.20, 2018, <https://github.com/illumina/pyflow/releases/tag/v1.1.20> [accessed: 31.08.2019].

7 Ibid.

8 Ibid.

The entire task management is automatically done by *pyFlow*. We just have to define the workflows in Python code and the engine handles all task dependencies and parallelization. Using *pyFlow* helps us to handle huge data inputs of several input files. The entire workload is automatically distributed among the maximum number of available CPU cores, which accelerates the entire process.

## Container virtualization with Docker

For our development and production environment we decided to deploy our Optical Character Recognition and natural language processing software in containers. For that purpose, we use a container virtualization software called *Docker*.<sup>9</sup> With Docker it is possible to easily create and use Linux containers.

“A Linux<sup>®</sup> container is a set of one or more processes that are isolated from the rest of the system. All the files necessary to run them are provided from a distinct image, meaning that Linux containers are portable and consistent as they move from development, to testing, and finally to production. This makes them much quicker than development pipelines that rely on replicating traditional testing environments. Because of their popularity and ease of use containers are also an important part of IT security.”<sup>10</sup>

In order to get a container up and running it is necessary to build a so called container image. An image is used to load a container in a predefined state at startup. Thus the image represents the initial state of a container including the chosen operating system base, software installations and configurations. It freezes a software deployment in its creation state and because of its portability it can then be shared easily.<sup>11</sup> That is why it is also suitable for publishing a software deployment in the context of a publication like the one at hand.

---

9 *Docker*, version 18.09.1 (Docker, 2013), <https://www.docker.com/> [accessed: 31.08.2019].

10 *Red Hat Inc.*, What's a Linux Container?, <https://www.redhat.com/en/topics/containers/whats-a-linux-container> [accessed: 20.05.2019].

11 *Boettiger, Carl*, An Introduction to Docker for Reproducible Research, in: ACM SIGOPS Operating Systems Review 49 (2015), 71–79, <https://doi.org/10.1145/2723872.2723882>.

To create a container image, Docker provides a build system which is based on so called *Dockerfiles*. These files act as a blueprint for the image the user wants to create. In order to create and maintain it, the developer only needs knowledge about operating a terminal and the concepts of the operating system used within the image.

Our images are based on the free and open source Linux distribution *Debian 9*.<sup>12</sup> We ensured that all our software that is installed on top of this basis is also free and open source software.

#### 4. A practical approach to optical character recognition of historical texts

As mentioned in the introduction, the first step of creating data is turning historical texts and sources into machine readable formats. The OCR process creates text files, namely XML, hOCR and PDF (with text layer). This process will be described in this part in detail. The corresponding source code and documentation can be downloaded from the GitLab page.<sup>13</sup> We mainly use the text files for further natural language processing steps which are described later in the text.

The entire process of turning books as well as other sources into data can be divided into a few manual pre- and postprocessing steps. The actual OCR is done automatically by our OCR pipeline.

Before the actual steps are described, we will briefly outline the goals of our pipeline and the software used. We also provide a short summary of the history of the OCR engine Tesseract.

##### Goals of our OCR pipeline: handling middle to large scale text input

This article proposes a practical way to do mid to large scale OCR for historic documents with a self developed Tesseract based pipeline. The pipeline is a tool to easily create mid to large sized text corpora for further research.

---

<sup>12</sup> *Debian*, version 9 (The Debian Project, 2017), <https://www.debian.org/> [accessed: 31.08.2019].

<sup>13</sup> The code for the OCR pipeline especially the pyFlow part is based on the original work of Madis Rumming, a former member of the INF team.

For now, the pipeline is a command line-based application which can be used to subject input documents to optical text recognition with a few simple commands. As of yet it is only used by the INF team of the SFB. Researchers have to request the OCR process for their sources and documents via our internal ticket system.

In the future, researchers will be able to upload any digitized TIFF or PDF document to the pipeline. It can handle multiple input documents, for example books, letters or protocols, at the same time and will automatically start the OCR process. Those documents will then be turned into text data. Researchers can choose between different languages per pipeline instance but not per document input.

To achieve this goal the INF team will build a virtual research environment (VRE).<sup>14</sup> The VRE will be implemented as a web application which can be easily used by every researcher of the SFB1288. Besides starting OCR processes researchers will also be able to start the tagging processes of text files. Different tagging sets will be available like the ones discussed in this article. Finally, researches will also be able to import tagged texts into an information retrieval system. We plan to either implement CQPweb<sup>15</sup> or use some of the provided application programming interfaces (APIs) to build our own front end.<sup>16</sup>

By providing this VRE the INF team will provide the researchers of the SFB1288 with many different tools which will aid them during different research steps.

## Implementation

All software dependencies needed to run our pipeline are documented in our source code repository hosted by the GitLab system of the Bielefeld University Library.<sup>17</sup>

---

14 The development of this VRE is in an early stage.

15 CQPweb is a web-based graphical user interface (GUI) for some elements of the IMS Open Corpus Workbench (CWB). The CWB uses the efficient query processor CQP, *Hardie, Andrew/Evert, Stefan*, IMS Open Corpus Workbench, <http://cwb.sourceforge.net/> [accessed: 31.08.2019].

16 *Sourceforge*, CWB/Perl & Other APIs, [http://cwb.sourceforge.net/doc\\_perl.php](http://cwb.sourceforge.net/doc_perl.php) [accessed: 13.05.2019].

17 The source code, documentation and pre-built images can be accessed here: <https://gitlab.uni-bielefeld.de/sfb1288inf/ocr> [accessed: 31.08.2019].

The pipeline consists of three files. The main file is *ocr* which implements the actual OCR pipeline. The file *parse\_hocr* is used to create a valid *DTA-Basisformat*<sup>18</sup> XML from the hOCR output files which follows the P5 guidelines of the Text Encoding Initiative (TEI).<sup>19</sup> The script *parse\_hocr* is called by the main file *ocr*. The last and third file is the Dockerfile used to automatically create an image. The image can then be used to start multiple containers to run multiple OCR processes. Detailed installation instructions can be found in the documentation.

## A short history of Tesseract

We use *Tesseract* for the actual OCR process. Tesseract is an open source OCR engine and a command line program. It was originally developed as a PhD research project at Hewlett-Packard (HP) Laboratories Bristol and at Hewlett-Packard Co, Greeley Colorado between 1985 and 1994.<sup>20</sup> In 2005 HP released Tesseract under an open source license. Since 2006 it has been developed by Google.<sup>21</sup>

The latest stable version is 4.0.0, which was released on October 29, 2018. This version features a new long short-term memory (LSTM)<sup>22</sup> network based

---

18 *Berlin-Brandenburgische Akademie der Wissenschaften*, Ziel und Fokus des DTA-Basisformats (Deutsches Textarchiv, Zentrum Sprache der Berlin-Brandenburgischen Akademie der Wissenschaften), [http://www.deutschestextarchiv.de/doku/basisformat/ziel.html#topic\\_ntb\\_5sd\\_qs\\_\\_rec](http://www.deutschestextarchiv.de/doku/basisformat/ziel.html#topic_ntb_5sd_qs__rec) [accessed: 12.03.2019].

19 *Text Encoding Initiative*, TEI P5: Guidelines for Electronic Text Encoding and Interchange (TEI Consortium) <https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf> [11.06.2019].

20 *Smith, Ray*, An Overview of the Tesseract OCR Engine, in: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2 (2007), <https://doi.org/10.1109/icdar.2007.4376991> [accessed: 31.08.2019].

21 *Google Inc.*, Tesseract OCR (2019), <https://github.com/tesseract-ocr/tesseract/> [accessed: 31.08.2019].

22 A special kind of recurrent networks, capable of using context sensitive information which is not near to the data which is processed (long-term dependencies). A LSTM network can be used to predict words in a text with respect of information which is further away. For example in a text it says that someone is from France and way later it says that this person speaks fluently x. Where x (= french) is the word to be guessed with the LSTM network by using the first information. *Olah, Christopher*, "Understanding LSTM Networks", <https://colah.github.io/posts/2015-08-Understanding-LSTMs/> [accessed: 26.03.2019].

OCR engine which is focused on line recognition.<sup>23</sup> Our own developed pipeline is based on the upcoming minor release, specifically on version 4.1.0-rc1 to benefit from the new neural net technology. During the development we also used version 3.05.01.

## Choosing data files for Tesseract

Tesseract needs models or so called data files per language for the OCR process. Models are manually trained. There are three main sets of trained data/data files for various languages available.

1. tessdata (Legacy models for Version 3)
2. tessdata\_fast (Fast standard models)
3. tessdata\_best (Slower for slightly better accuracy)

For now we are using the following models for the corresponding languages:

- German model from tessdata\_best<sup>24</sup>
- German Fraktur not available as tessdata\_best<sup>25</sup>
- English model from tessdata\_best<sup>26</sup>
- English middle from tessdata\_best<sup>27</sup>
- French from tessdata\_best<sup>28</sup>
- French middle from tessdata\_best<sup>29</sup>
- Portuguese from tessdata\_best<sup>30</sup>
- Spanish from tessdata\_best<sup>31</sup>

---

23 Google Inc., Tesseract OCR.

24 [https://github.com/tesseract-ocr/tessdata\\_best/raw/master/deu.traineddata](https://github.com/tesseract-ocr/tessdata_best/raw/master/deu.traineddata)

25 [https://github.com/tesseract-ocr/tessdata/raw/master/deu\\_frak.traineddata](https://github.com/tesseract-ocr/tessdata/raw/master/deu_frak.traineddata)

26 [https://github.com/tesseract-ocr/tessdata\\_best/raw/master/eng.traineddata](https://github.com/tesseract-ocr/tessdata_best/raw/master/eng.traineddata)

27 [https://github.com/tesseract-ocr/tessdata\\_best/raw/master/enm.traineddata](https://github.com/tesseract-ocr/tessdata_best/raw/master/enm.traineddata)

28 [https://github.com/tesseract-ocr/tessdata\\_best/raw/master/fra.traineddata](https://github.com/tesseract-ocr/tessdata_best/raw/master/fra.traineddata)

29 [https://github.com/tesseract-ocr/tessdata\\_best/raw/master/frm.traineddata](https://github.com/tesseract-ocr/tessdata_best/raw/master/frm.traineddata)

30 [https://github.com/tesseract-ocr/tessdata\\_best/raw/master/por.traineddata](https://github.com/tesseract-ocr/tessdata_best/raw/master/por.traineddata)

31 [https://github.com/tesseract-ocr/tessdata\\_best/raw/master/spa.traineddata](https://github.com/tesseract-ocr/tessdata_best/raw/master/spa.traineddata)

We aim to only use trained data from `tessdata_best` to achieve high quality OCR results. Only German Fraktur is not available as a `tessdata_best` model.

## Overview of the entire OCR process and the pipeline

This part describes the function of the developed pipeline in detail, beginning with the digitization and preprocessing of input documents. Following these steps, the actual process of OCR is described in general to provide an overview of the underlying principles and technologies used. Lastly, the output files of the pipeline are described, and we explain why those files are generated and what they are used for.

In addition to this we also discuss the accuracy of the OCR and how it can affect the text data output as well as further research using the data.

Figure 3 shows the entire OCR process including manual and automatic steps. Every step is discussed in the following parts.

## Input for the pipeline: digitization and collection of historic documents

As mentioned above the pipeline accepts TIFF (only multi-page TIFFs per document) and PDF files as input. These input files have to be obtained or created first. The process of creating input files as discrete sets of pixels from physical media like paper based books, etc. is called digitization.<sup>32</sup> Scanning a book and creating a PDF file from it is only one example of digitization, however. Another example is taking a picture of a document with the camera of a mobile phone, which creates a JPEG file. Both examples obviously result in digital representations of different quality.

Digitization is the first step of the full OCR process as depicted in figure 3.

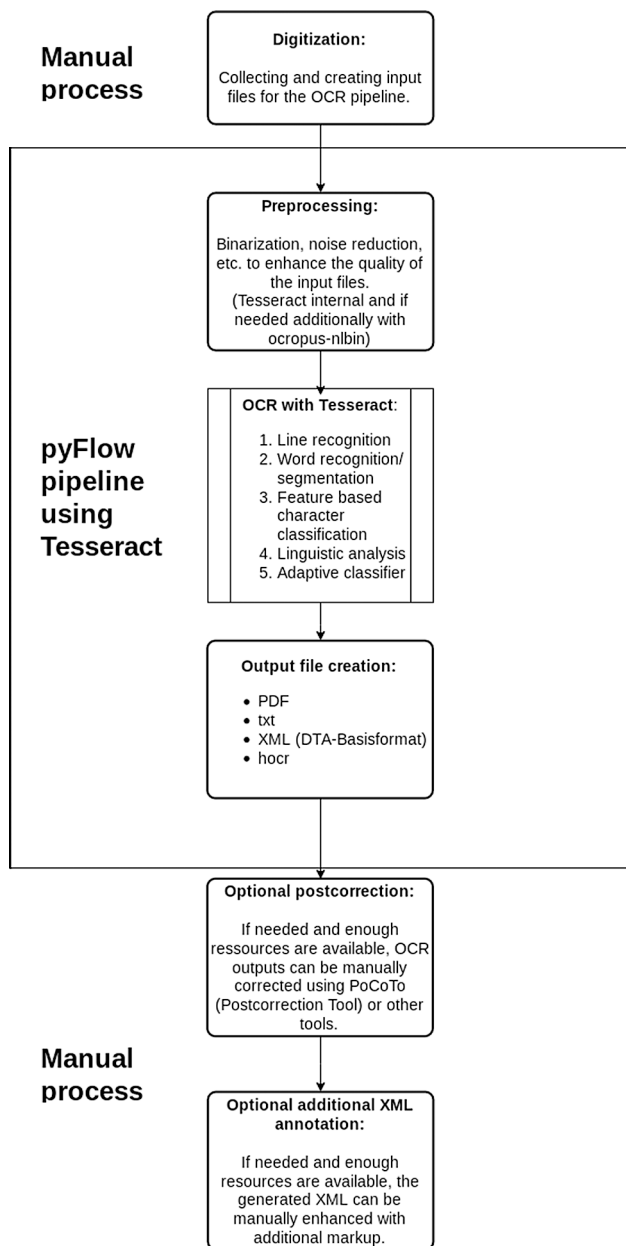
Below we describe the most common ways how researchers of the SFB are obtaining or creating input files for the pipeline. The different ways result in different qualities of input files.

---

32 Ye, Peng/Doermann, David, Document Image Quality Assessment: A Brief Survey, in: 2013 12th International Conference on Document Analysis and Recognition (2013), 723, <https://doi.org/10.1109/icdar.2013.148>.



Fig. 3: The entire OCR process



The first and most common way to obtain input files for the pipeline is to use digitized documents provided by libraries or other institutions. These files vary in quality depending on how they were created.

It is also common to obtain image files from libraries which were created from microfilms. This way is similar to obtaining scans of books like mentioned above. Unfortunately, sometimes those images are of poor quality.

Our experience has shown that images obtained from libraries do not meet our quality demands sometimes. Whenever this is the case and a physical copy of said document is at hand, we repeat the digitization within our own quality parameters.

We advise to always assess the quality of input images based on the criteria listed below. Especially background noise and geometric deformation decisively decrease the quality of the OCR process.

It may sound as if libraries in general do not do a very good job of creating high quality digital representations of books and other documents. This is not the case because we have to keep in mind that the digitization is mainly done with human readers in mind. Humans are far better at reading documents of relatively low quality than computers. The demand for high quality images used for OCR processes and thus for corpus linguistic projects has risen over the years. Most of the libraries are adapting to this new trend and are providing the necessary images.

The third way to obtain input files is to perform our own scans of books and microfilms. This should always be the method of choice if already obtained digitized input files are of low quality. In general, it is better to always perform own scans with predefined parameters to ensure the best possible quality.

During the stage of obtaining and creating input files we can already enhance the accuracy of the subsequent OCR process significantly. It should always be the goal to obtain or create image files of the highest quality. The higher the quality of the scans the better are the end results of the OCR, which ultimately leads to high quality text data corpora.

There are a few criteria on how to determine if a digital representation is of good quality:<sup>33</sup>

- 1) Stroke level
  - a) Touching characters
  - b) Broken characters
  - c) Additive noise:
    - i) Small speckle close to text (For example dirt.)
    - ii) Irregular binarization patterns
- 2) Line level
  - a) Touching lines
  - b) Skewed or curved lines
  - c) line inconsistency
- 3) Page level
  - a) background noise:
    - i) Margin noise
    - ii) Salt-and-pepper
    - iii) Ruled line
    - iv) Clutter
    - v) Show through & bleed through
    - vi) Complex background binarization patterns
  - b) Geometric deformation:
    - i) Warping
    - ii) Curling
    - iii) Skew
    - iv) Translation
- 4) Compression methods
  - a) Lossless compression methods are preferred

When creating our own scans, we can follow some best practices outlined below to avoid some of the above mentioned problems with the digital representation of documents:

---

33 P. Ye/D. Doermann, Document Image Quality Assessment: A Brief Survey, 724.

General best practices are mainly based on the official Tesseract wiki entry:<sup>34</sup>

- Create scans only with image scanners. Other digitization measures like using mobile phone cameras etc. are not advised. These will easily introduce geometric deformation, compression artifacts and other unwanted problems all mentioned in this list and the criteria list above. We mention this especially because we often had to handle images of books that have been taken with mobile phone cameras or other cameras.
- Avoid creating geometric deformation like skewing and rotation of the page. (This can be hard with thick books because the book fold will always introduce some warping.)
- Avoid dark borders around the actual page. (These will be falsely interpreted as characters by Tesseract.)
- Nonetheless also avoid not having any border or margin around the text.
- Avoid noise like show through, bleed through etc.
- One page per image recommended. Do not use double-sided pages.

Technical specifications also mainly based on the official Tesseract wiki entry:<sup>35</sup>

1. Scan with 300 dots per inch (DPI)
2. Use lossless compression
  - a) Avoid JPEG compression methods. This method introduces compression artifacts and compression noise around the characters.<sup>36</sup> This is due to the discrete cosine transform (DCT) which is part of the JPEG compression method.<sup>37</sup> (Figure 4 shows an example of the differences between a lossless compressed TIFF and a lossy compressed jpeg.)
  - b) Avoid lossy compression methods in general.

---

34 Google Inc., ImproveQuality: Improving the Quality of the Output (2019), <https://github.com/tesseract-ocr/tesseract/wiki/ImproveQuality> [accessed: 31.08.2019].

35 Ibid.

36 P. Ye/D. Doermann, Document Image Quality Assessment, 723.

37 Oztan, Bazak, et al., Removal of Artifacts from JPEG Compressed Document Images, in: Reiner Eschbach/Gabriel G. Marcu (eds.), Color Imaging XII: Processing, Hardcopy, and Applications, (SPIE) 2007, 1-3, <https://doi.org/10.1117/12.705414>.

- c) This is why we use and recommend TIFF files with lossless compression using the Lempel-Ziv-Welch-Algorithm (LZW-Algorithm or LZW)<sup>38</sup>
- d) We also accept PDFs. (We have to admit that this is a trade off for convenience. PDFs can be of the same quality as TIFF files if they are created from images using the FLATE/LZW compression. The default though is lossy JPEG compression.)<sup>39</sup>

*Fig. 4: Lossless and lossy image compression*



The top string shows a lossless compressed TIFF file. The lower string shows a lossy compressed JPEG file. Both files are binarized. The JPEG compression rate is 70 to give a better visual example. (self-created)

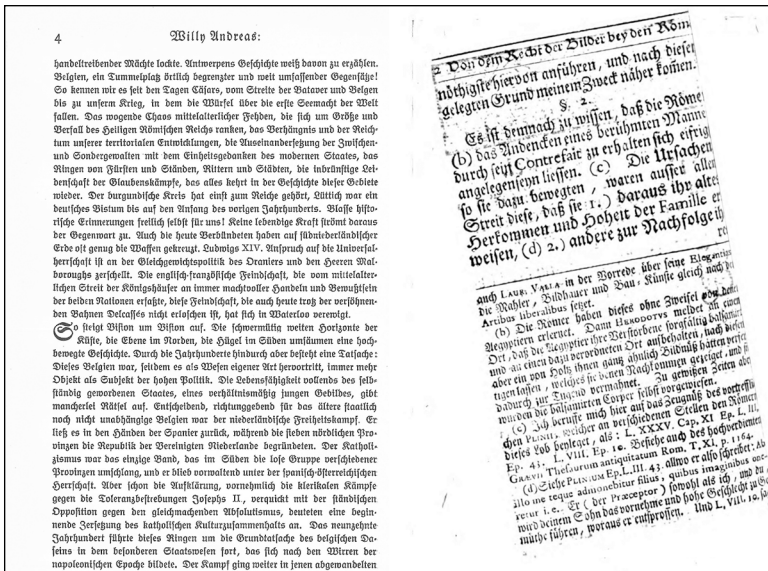
To sum up, digitization of historical documents is already an important step significantly influencing the accuracy of the OCR process. In this part we have outlined how we obtain and create input files for our OCR pipeline. The main goal is to always obtain or create input files of the highest quality as possible. To achieve this, we described criteria to determine the quality of given input files. Additionally, we described some rules on how to create high quality scans of historic documents.

---

38 *Adobe Systems Incorporated*, TIFF: Revision 6.0, version 6.0 (1992), 57–58, <https://www.adobe.io/content/dam/udp/en/open/standards/tiff/TIFF6.pdf> [accessed: 31.08.2019].

39 *Adobe Systems Incorporated*, Document Management – Portable Document Format – Part 1: PDF 1.7 (2008), 25 ff., [https://www.adobe.com/content/dam/acom/en/devnet/pdf/PDF32000\\_2008.pdf](https://www.adobe.com/content/dam/acom/en/devnet/pdf/PDF32000_2008.pdf) [accessed: 26.03.2019].

Fig. 5: Comparison between high and low quality



The left part of the figure shows a high quality scan. The right part shows a low quality scan with several problems like background noise and geometric deformation. (self-created)

## Starting the pipeline with input files and user set parameters

In the next step the collected and created files are passed into the actual OCR pipeline. The OCR pipeline is written in python using the pyFlow package. It is deployed by using Docker. A description of Docker and pyFlow can be found in part 3.

As soon as this is done, we can place our input files in the folder “files\_for\_ocr”, which was created during the setup of the pipeline. Files must be sorted into folders beforehand on a per document basis. For example, a scanned book results in one multi-page TIFF file. This file has to be placed into its own corresponding folder inside the folder *files\_for\_ocr*. PDFs should be handled the same way. We can put as many folders into the input folder of the pipeline as we want. The only constraint here is that every input document should be of the same language because we tell Tesseract to use a specific language model for the OCR process. Because the pipeline is written using pyFlow it creates different processes for each document in the input folder and works

through those in an efficient manner (see also part 3.). To handle multiple documents of different languages at the same time we recommend starting a new Docker instance per language and place the documents per language in the corresponding input folder of the Docker instance. If a document consists of multiple languages it is possible to tell Tesseract which language models it should use at the same time.

Once every input file has been put into the folder *files\_for\_ocr* we can start the pipeline with a simple command. The command and further examples can be found in our corresponding GitLab repository documentation.<sup>40</sup>

The following parameters can be set by the user:

### *Language*

This parameter tells Tesseract which model it should use for the OCR process. Language should be set to the corresponding language of the input files.

### *Binarization*

The user can decide to binarize the input images with *ocropus-nlbin* in an additional upstream preprocessing step (see below for more information on binarization).

## **Pipeline processing step 1: unify input files**

Before the preprocessing of the input files starts, the pipeline converts PDF files into TIFF files using the package *pdftoppm*. Every page of the PDF is converted into one TIFF file with 300 DPI using the lossless LZW compression method. Multi-page TIFF files are split per page into individual files. Now that all input files are of the same type every following preprocessing step (e. g., binarization) can be applied to those uniformly file by file.

## **Pipeline processing step 2: preprocessing of the input files**

The first internal step of the pipeline is preprocessing of the input images. Some major steps in enhancing the image quality are described below.

---

<sup>40</sup> Jentsch, Patrick/Porada, Stephan, Docker Image: Optical Character Recognition, version 1.0, Bielefeld 2019, [https://gitlab.uni-bielefeld.de/sfb1288inf/ocr/container\\_registry](https://gitlab.uni-bielefeld.de/sfb1288inf/ocr/container_registry) [accessed: 31.08.2019].

## Binarization and noise removal

One important first step in preprocessing is binarization. The goal of binarization is to reduce the amount of noise and useless information in the input image.<sup>41</sup>

In general binarization is the step of converting a color image into a black and white image. The idea is to only extract the pixels which actually belong to the characters and discard any other pixel information which, for example, is part of the background. To achieve this the technique of thresholding is used. Basically, the method of thresholding analyses each pixel of a given picture and compares its grey level or another feature to a reference value. If the pixels value is below the threshold it will be marked as a black pixel and thus as belonging to a character. If the value of the pixel is above the threshold it will be labeled as white pixel und thus be handled as the background. Binarization techniques using thresholding can be divided into two classes: global and local thresholding. Both methods differ in what reference value for the pixel comparison is being used. Global thresholding calculates one reference value per pixel in one picture while local thresholding calculates the reference value for each pixel based on the neighboring pixels.<sup>42</sup> Figure 6 shows the successfully applied binarization step done with `ocropus-nlbin`.

The pipeline will always use the built in Tesseract binarization. Tesseract's built in binarization uses the Otsu algorithm.<sup>43</sup> There is also the option to binarize the pictures before passing them to Tesseract, if the built in binarization is not sufficient. For this additional upstream binarization process our pipeline uses the `ocropus-nlbin` library. This step can easily be invoked by using the corresponding parameter (see our documentation).

Omitting or explicitly using this additional upstream step can in both cases either result in better or worse accuracy. Which option is chosen is

---

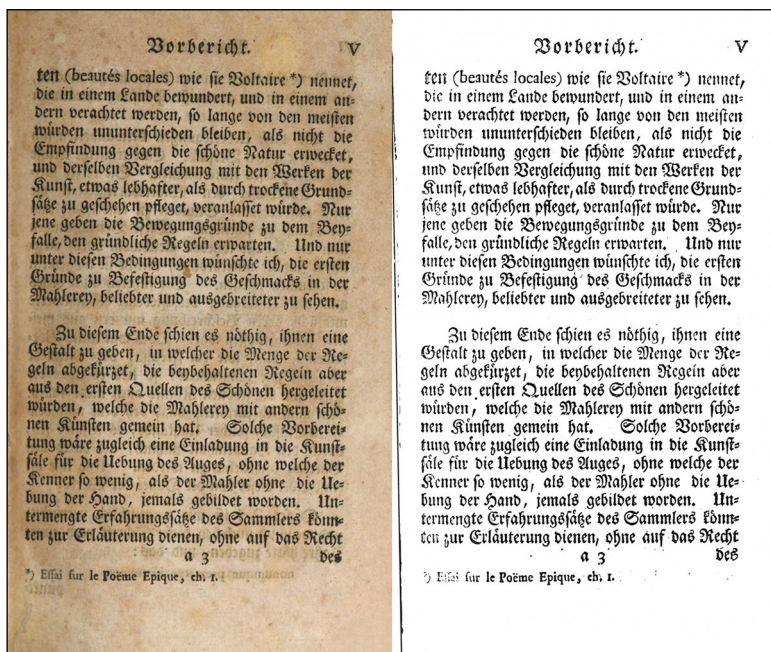
41 Chaudhuri, Arindam, et al., *Optical Character Recognition Systems for Different Languages with Soft Computing*, Springer International Publishing: 2017, 90–92, 17–22, <https://doi.org/10.1007/978-3-319-50252-6>.

42 Cheriet, Mohamed, et al., *Character Recognition Systems*, John Wiley & Sons, Inc.: 2007, 8–15, <https://doi.org/10.1002/9780470176535>.

43 Google Inc., `ImproveQuality`: Improving the Quality of the Output.



Fig. 6: Binarization process



The left part shows an input image before the binarization step has been applied. The right part shows the same image after the binarization step. Note that during the binarization minimal skew is also automatically removed. (self-created)

decided on a per input file basis taking the input files quality into account. Sometimes we also evaluate the output accuracy of one step and rerun the OCR process to achieve a better accuracy by using or not using ocropus-ml-bin. Subchapter Accuracy in Part 5 discusses the different accuracy values for the use and non-use of the additional upstream binarization step for some example files.

## Skew detection and correction

Even when using scanners for the digitization process, there will always be a few degrees of tilt or skew of the text due to human influence (for example how the book was placed inside the scanner etc.).

Tesseract's line finding algorithm was designed to work without having to deskew a page to successfully identify text lines. This design choice avoids the loss of image quality.<sup>44,45</sup> Manual deskewing is only needed if the skew is too severe, as mentioned in the official wiki.<sup>46</sup> According to the original paper from 1995 describing the algorithm<sup>47</sup> the line finding algorithm produces robust results for angles under 15 degrees. Because of this we have not implemented automatic deskewing of input files.

If the skew of the input files is too severe, they have to be deskewed manually before passing them into the pipeline. Manual deskewing is also advised because the skew can vary severely from page to page. Automatic deskewing could therefore result into the loss of text parts, depending on the discrepancies in skew between pages.

After the preprocessing steps, the actual OCR process starts. The process is described in the following part.

### Pipeline processing step 3: OCR process

This part gives an overview of the internal steps of the actual OCR process. We describe which steps are performed by Tesseract internally to perform the OCR process. Some of the steps are done by various other OCR engines in general, some of the steps are specific to Tesseract. Those differences are highlighted and explained.

One of the first steps performed by Tesseract is line finding.<sup>48</sup> This step is specific to Tesseract because the algorithm was explicitly designed for it.<sup>49</sup> In general this step detects lines of text in already provided and identified text regions. One advantage of the algorithm is its achievement of robust results for line recognition on pages with a skew of up to 15 degrees.

---

44 R. Smith, An Overview of the Tesseract OCR Engine.

45 Smith, Ray, A Simple and Efficient Skew Detection Algorithm via Text Row Accumulation, in: Proceedings of 3rd International Conference on Document Analysis and Recognition (IEEE Comput. Soc. Press, 1995), <https://doi.org/10.1109/icdar.1995.602124>.

46 Google Inc., ImproveQuality: Improving the Quality of the Output.

47 R. Smith, A Simple and Efficient Skew Detection Algorithm via Text Row Accumulation.

48 R. Smith, An Overview of the Tesseract OCR Engine.

49 R. Smith, A Simple and Efficient Skew Detection Algorithm via Text Row Accumulation.

Once the lines have been identified, additional steps are being executed to fit the baseline of every line more precisely. Tesseract handles curved lines especially well.<sup>50</sup> This is another advantage of the algorithm as curved lines are a common artifact in scanned documents.

The next major step executed by Tesseract is word recognition. This step is also performed by various other available OCR engines. The goal of word recognition is to identify how a word should be segmented into characters.<sup>51,52</sup> For this step characters have to be recognized and then chopped or segmented.

Another part of the Tesseract OCR process is the so called *Static Character Classifier* or, in more general terms, the character classification. This step is feature based which is a common approach in various available OCR engines.<sup>53,54</sup> The goal of feature extraction is to identify essential characteristics of characters.<sup>55</sup> Based on the extracted features the classification process will be executed. Each character will be identified based on its features and will be assigned to its corresponding character class.<sup>56,57</sup>

One of the last steps Tesseract executes is a linguistic analysis. It only uses a minimal amount auf linguistic analysis.<sup>58</sup> Tesseract, for example, compares every word segmentation with the corresponding top dictionary word. This process is a statistical approach where the segmentation will be matched with the most likely corresponding word. This process is done for other categories besides the top dictionary word.<sup>59</sup>

These are the main steps done by Tesseract to turn images into machine readable text.

---

50 R. Smith, An Overview of the Tesseract OCR Engine.

51 M. Cheriet et al., Character Recognition Systems, 204–206.

52 R. Smith, An Overview of the Tesseract OCR Engine.

53 Ibid.

54 A. Chaudhuri et al., Optical Character Recognition Systems for Different Languages with Soft Computing, 28.

55 Ibid.

56 A. Chaudhuri et al., Optical Character Recognition Systems for Different Languages with Soft Computing, 28.

57 R. Smith, An Overview of the Tesseract OCR Engine.

58 Ibid.

59 Ibid.

## Pipeline processing step 4: output file creation

Tesseract can automatically create several output files after the OCR process is finished. Output files will be created per TIFF file. The file formats are:

### *hOCR*

Standard output file. HTML based representation of the recognized text. Includes position of lines and characters matching the corresponding input image. Can for example be used to create PDFs with an image layer using the original input image. Also needed for post correction with PoCoTo. (See Pipeline processing step 5 in part 4.)

### *PDF*

Tesseract automatically creates one PDF file per page consisting of an image layer and an invisible text layer. The image layer shows the input TIFF file. The text layer is placed in such a way that the recognized strings match the actual visible text in the image.

Besides those two outputs the pipeline automatically creates the following files per input document:

- Combined PDF (combines the single PDF pages into one file per document)
- Combined text file (created from the combined PDF file)
- DTA-Basisformat XML (created from the hOCR files per document)

The combined PDF file is created from the single page PDFs containing the image and text layer. The combined PDF files are mainly created for humans because they are easily readable on any device.

The text file per document is created from the text output files of Tesseract. For that purpose we use a simple bash command which utilizes `cat`. We aim to export paragraphs and other formatting structures. The text files are mainly used as input for further computer linguistic methods. These methods are described in part 5.

Lastly, the pipeline automatically creates valid DTA-Basisformat XML files per input document. The XML files are created from the hOCR output files. The DTA-Basisformat XML structure follows the P5 guidelines of the

Text Encoding Initiative (TEI).<sup>60</sup> The DTA-Basisformat is developed by the Deutsches Textarchiv and recommended for digitizing and archiving historical texts.<sup>61</sup> One goal of the XML markup is to preserve the logical structure of the digitized texts. For example, headings, paragraphs and line breaks are being annotated with corresponding tags. Also, procedural markup of text color or italic written text is being annotated. The DTA-Basisformat syntax can also be used to annotate more uncommon text parts like poems, recipes, marginal notes or footnotes.

Creating the output files is the last automatic step done by the pipeline. Files can now be enhanced and corrected during postprocessing steps or be passed to the next process (POS tagging, NER etc.).

### **Pipeline processing step 5: optional manual postprocessing steps**

If needed, manual postcorrection of the output texts can be done. For this, the post correction tool PoCoTo can be used. With this it is possible to use the hOCR files in conjunction with the corresponding TIFF files. PoCoTo gives the user a side by side view where one can compare the recognized text with the actual image representation. If the text does not match the image, the user can correct it. It is also possible to correct common repeated errors automatically and thus save time.

Alternatively, every common text editor can be used to correct the text in the hOCR files directly.

After the post correction, new PDFs, XML and text files have to be created manually.

Besides a simple postcorrection of the text, another manual and optional step would be the enhancement of the XML markup. Tesseract and our pipeline recognize paragraphs and line breaks, which are automatically written into the XML file. More sophisticated elements must be annotated by hand. The annotation can be done with any common text editor. Common elements that have to be annotated manually are marginal notes or footnotes.

---

<sup>60</sup> *Text Encoding Initiative*, TEI P5: Guidelines for Electronic Text Encoding and Interchange.

<sup>61</sup> *Berlin-Brandenburgische Akademie der Wissenschaften*, Ziel und Fokus des DTA-Basisformats.

## Evaluation accuracy of Tesseract

In this part we talk about the accuracy and error rates of Tesseract and our pipeline. First, we briefly show some error rates and accuracy values for the Tesseract OCR engine published by Ray Smith working for Google.

Besides the official numbers we also show some results from our own accuracy tests performed with collected and self created test data as an input for our pipeline. The test data consists of digital input images and the corresponding manually created and corrected accurate textual content of those. In the context of OCR this transcription is called ground truth data.<sup>62</sup>

For these accuracy tests we will input the test data images into the pipeline and compare the output text with the ground truth text. From the discrepancies between the output and the ground truth data we can calculate two different error rates.

In the context of OCR evaluation two metrics are used to describe the error rate at two different levels: Character error rate (CER) and Word error rate (WER). Both metrics are calculated independently in regard to the length of the output text data. In order to achieve this, the number of mistakes is divided by the text length resulting in an error rate either for words or characters. This has to be done for both metrics.<sup>63</sup>

For the concrete calculation of those error rates we use the *OcrevalUAtion* tool.<sup>64,65</sup> This tool compares the ground truth text with the actual output text from the OCR pipeline and calculates the error rates accordingly.

The goal of our own accuracy tests is to see if our digitization, preprocessing and binarization steps are either beneficial or disadvantageous for the accuracy of the OCR process. We also want to measure the quality difference between files that have been binarized with *ocropus-nlbin* and those

---

<sup>62</sup> Carrasco, Rafael C., Text Digitisation, <https://sites.google.com/site/textdigitisation/> [accessed: 11.06.2019], ch. 2.1.

<sup>63</sup> Ibid.

<sup>64</sup> University of Alicante, *ocrevalUAtion*, <https://github.com/impactcentre/ocrevalUAtion> [accessed: 05.04.2019].

<sup>65</sup> University of Alicante, *OcrevalUAtion*, version 1.3.4 (2018), <https://bintray.com/impactocr/maven/ocrevalUAtion> [accessed: 31.08.2019].

that have not. In addition to that we can also compare our values to the ones published by Google.

We focus on the values for modern English and Fraktur.

## Official numbers

The published accuracy results for Tesseract show quite low error rates for Latin languages like English and French. Tesseract version 4.0+ using the LSTM model and an associated dictionary check has an CER of 1.76 for English text. The WER is 5.77. The CER for French is 2.98 and the WER is 10.47. In general, the error rates for Latin languages are in a similar range.<sup>66</sup> Note that these low error rates are probably the result of high quality image inputs. We can deduce this from our own calculated error rates shown in the following part.

## Own tests with ground truth data

Table 1 shows our own calculated error rates for different input data. We tested our pipeline with input images of two different quality levels. High quality images are TIFFs we created with our own scanners. These images were created with a minimum of 300 DPI and full color range. We also made sure that we introduced as little skew and rotation as possible during the scanning process. Input images of medium quality were created from PDF files with lower DPI and possible JPEG compression artifacts. Skew and rotation levels are still minor though. Ground truth data exists for every file, either self created from OCR with postcorrection or from available online sources. Images, ground truth data and the accuracy test results can be found in detail in the respective GitLab repository.<sup>67</sup>

---

<sup>66</sup> *Smith, Ray*, Building a Multilingual OCR Engine: Training LSTM Networks on 100 Languages and Test Results (Google Inc., June 20, 2016), 16, 17, [https://github.com/tesseract-ocr/docs/blob/master/das\\_tutorial2016/7Building%20a%20Multi-Lingual%20OCR%20Engine.pdf](https://github.com/tesseract-ocr/docs/blob/master/das_tutorial2016/7Building%20a%20Multi-Lingual%20OCR%20Engine.pdf).

<sup>67</sup> [https://gitlab.uni-bielefeld.de/sfb1288inf/ground\\_truth\\_test](https://gitlab.uni-bielefeld.de/sfb1288inf/ground_truth_test).

Table 1: Accuracy test results (self-created)

	Quality	Quality features	Document	Pages	CER <sup>a</sup>	WER <sup>b</sup>	CER <sup>c</sup>	WER <sup>d</sup>
Fraktur	High	300 DPI, self created TIFF scans	Estor - Rechts- gelehrsamkeit	4	19.45	45.58	23.61	58.78
			Luz - Blitz	4	19.91	44.54	22.28	61.76
	Middle	TIFFS created from PDFs	Die Gegenwart	10	12.07	19.71	5.38	10.42
English	High	300 DPI, self created TIFF scans	Inside Germany	10	2.60	5.86	1.62	1.88
			Germans Past and Present	10	4.17	5.69	4.00	5.13

a Additional binarization with Ocropus.

b Additional binarization with Ocropus.

c Only internal Tesseract binarization.

d Only internal Tesseract binarization.

We tested the OCR pipeline with every input document twice, each time using different parameters. The first run utilized the additional binarization step using ocropus-nlbin from ocopy. After the OCR process had finished, we calculated CER and WER with the ocrevalUation tool. For the second test run we omitted the additional binarization step and only used the internal binarization step provided by Tesseract. CER and WER were calculated accordingly.

As we can see the pipeline achieves low error rates for Modern English high quality input images. CER is 2.6 and WER is 5.86. When not using the additional binarization step the results are even better with CER being 1.62 and WER being 1.88. If we compare those findings to the results published by Ray Smith, we can see that we achieved slightly better results. We attribute this to the high quality of our self created input images. They were created in accordance with our own best practices, as outlined above.

Error rates for English medium quality input images are slightly worse but still in close range to the error rates of high quality input images.



In general, the error rates for Fraktur text are much higher than for English text. This could be linked to several factors. First, as mentioned above, the model for German Fraktur is not available from the `tessdata_best` set. Second, because of their age, Fraktur texts are more often susceptible to background noise, touching or broken characters and geometric formation. This could also explain the higher error rates for supposedly high quality Fraktur input images compared to medium quality input images. On paper the high quality input images have a much higher DPI but suffer more from bleed through, line skew and broken characters (fading characters), etc. than the middle quality input images.

From these findings we can conclude that the quality level of the input images should be seen as a two dimensional parameter consisting of technical quality (DPI, lossless compression, etc.) and physical quality of the text (fading characters, skew, bleed through, etc.).

Regarding the additional binarization step, it is hard to judge when it is beneficial. For the OCR of Fraktur text the results suggest that it could be beneficial in some cases. Possible researchers should run the OCR process twice, once with additional binarization and once without it, and judge for themselves which output has fewer errors.

For English text the results suggest that additional binarization is not beneficial.

To give a finite answer we would have to do more testing with more diverse ground truth data.

## Natural language processing

By using natural language processing (NLP) methods it is possible to enrich plain texts with various useful information. Our goal after processing is to make the source searchable for the added data. For that purpose, we decided to use the free open source NLP library `spaCy`. It is fast, reliable and offers natural language processing for all languages used in our context to the same extent. The latter is not self-evident, other open source approaches we have tried, like *Stanford CoreNLP*, do not provide all features we want to make use of for all languages.

It was important that we were able to handle each text corpus in the same way, independently of the input language. For further work with the gathered data we use the software collection *The IMS Open Corpus Work-*

*bench* (CWB) which uses a data type called *verticalized text* (*vrt*) format. This data type is a fairly uncommon variation of XML, which is why most of the NLP libraries do not offer it as an output option. Because we did not want to perform much data type conversion, we needed a flexible NLP toolkit with which we could configure the output format with an, in the best case, application programming interface (API).

For the time being it is enough for us to use four methods for further text analysis. These are tokenization, lemmatization, part-of-speech tagging (POS tagging) and named entity recognition, which are all described in the following. It is good to know that our chosen natural language processing toolkit offers this and gives us the possibility to extend this portfolio with more features in the future.

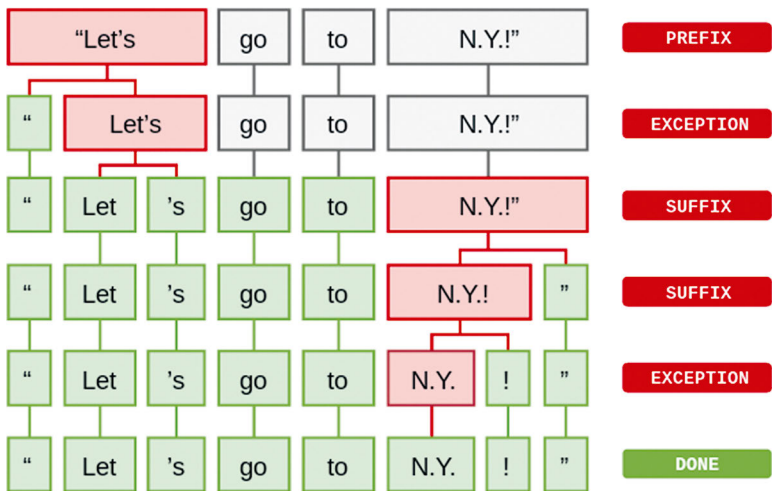
After performing NLP on a text, it is possible to fulfill queries, like “show all text passages where words of the category adjective appear around the word ‘world’ in all its reflections within a maximum word spacing of 5 words”, which can help you by finding assessments made of the world in the queried text.

## 5. Tokenization

Tokenization is the task of splitting up a text into so called tokens. These can either be words or punctuation marks.<sup>68</sup> This process can be easily explained by a short example. The sentence “Let’s go to N.Y.!” will get tokenized into eight tokens: “, Let, ‘s, go, to, N.Y., !, ”. As you can see, tokenization is not just about splitting by white space or non-alphanumeric characters. The tokenizer system needs to recognize that the dots between and after the “N” and “Y” do not indicate the end of a sentence, and that “Let’s” should not result in one token but in two: “Let” and “’s”. The process of tokenizing differs a bit in each software implementation. The spaCy tokenizer, which is used by us, is shown in figure 7.

---

68 Manning, Christopher D./Raghavan, Prabhakar/Schutze, Hinrich, Introduction to Information Retrieval, Cambridge: Cambridge University Press, 2008, 22, <https://doi.org/10.1017/cbo9780511809071>.

Fig. 7: Tokenization process with *spaCy*

First, the tokenizer splits the sentence into tokens by white space. After that, an iterative process starts in which the tokenizer loops over the gained tokens until the process is not interrupted by exceptions anymore. An exception occurs when tokenization rules exist that can be applied to a token.<sup>69</sup>

The tokenization rules which trigger the exceptions are usually extremely specific to the language used in the text. *SpaCy* offers the ability to extend its features with predefined language packages for all languages we want to support. These packages already include the language specifics. They are easily expandable by adding your own rules for special expressions in texts. The latter can be the case if your text uses an old style of a language – which is likely to appear in historical texts – which is not processable by modern tokenization rules that are designed for today's language compositions.

<sup>69</sup> Image from *Explosion AI*, Linguistic Features · *spaCy* Usage Documentation: Tokenization, <https://spacy.io/usage/linguistic-features#tokenization> [accessed: 25.03.2019].

## Lemmatization

The grammar, which forms a language, leads to different forms of a single word. In many situations where you want to use methods for text analysis, you are not interested in a specific form but in the occurrence of any form. To make it possible to query for those occurrences we use a lemmatization process, which adds a lemmatized version to each token of a text. The lemmatized version of a word is defined as the basic form of it, like the one which can be found in a dictionary.<sup>70</sup> As an explanatory example the following words are lemmatized like this:

- writes, wrote, written → write
- am, are, is → be
- car, cars, car's, cars' → car

The lemmatization process is based on a dictionary lookup where a lemma dictionary (basically a long list of words in all its inflections with a corresponding lemmatized entry) is used to lemmatize words. These dictionaries are usually bundled in the NLP toolkit's language packages. If the software does not find a word which should be lemmatized in the corresponding dictionary, it just returns the word as it is. This is a pretty naive implementation, which is why for some languages the developers added some generic rules which are applied after a dictionary lookup fails.

For our chosen NLP toolkit, the lemmatization implementations can be found in the source code repositories of Explosion AI. These repositories show why their English lemmatizer performs better than most of their others like, for example, the German one. It is not only based on dictionaries but also on more generic rules specific to the language.<sup>71,72</sup>

---

<sup>70</sup> C. D. Manning/P. Raghavan/H. Schütze, Introduction to Information Retrieval, 32.

<sup>71</sup> Explosion AI, Industrial-Strength Natural Language Processing (NLP) with Python and Cython: SpaCy/Spacy/Lang/de/Lemmatizer.py, <https://github.com/explosion/spaCy/tree/v2.1.0/spacy/lang/de/lemmatizer.py> [accessed: 21.05.2019].

<sup>72</sup> Ibid.

## Part-of-speech tagging

Words can get categorized into different classes. Knowing the class of a word is useful because they offer further information about the word itself and its neighbors. Part-of-speech tagging is exactly about that: Tokenized texts are analyzed, and a category of a predefined set is assigned to each token.<sup>73</sup> These category sets are also called ‘tagsets’. SpaCy’s POS tagging process is based on a statistical model which improves the accuracy of categorization predictions by using context related information, for example, a word following “the” in English is most likely a noun.<sup>74</sup> The part-of-speech tagsets used by spaCy are based on the chosen language model.<sup>75</sup>

## Named entity recognition

Named entity recognition (NER) is the task of detecting named entities in a text. These can be understood as anything that can be referred to with a name, like organizations, persons and locations. Named entities are quite often ambiguous, the token `Washington`, for example, can refer to a person, location, and an organization.<sup>76</sup> Our used NLP software offers a system that detects named entities automatically and assigns a NER-tag to the corresponding token. The release details of spaCy’s language packages list all NER-tags that are assigned by it.

## Accuracy

Having NLP software that satisfies your needs in terms of functionality is important, nevertheless you should be aware of its reliability. Table 2 gives an overview of the accuracies of spaCy’s language packages.

---

73 Jurafsky, Daniel/Martin, James H., *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Draft of September 23, 2018, 2018, 151, 156.

74 *Explosion AI*, *Linguistic Features spaCy Usage Documentation: Tokenization*.

75 *Explosion AI*, *Annotation Specifications · spaCy API Documentation: Part-of-Speech Tagging*, <https://spacy.io/api/annotation#pos-tagging> [accessed: 27.03.2019].

76 D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 328–29.

Table 2: Accuracy values of spaCy

	NER	POS tagging	Tokenization
Dutch <sup>a</sup>	100.00	87.02	91.57
English <sup>b</sup>	99.07	86.56	96.92
French <sup>c</sup>	100.00	82.64	94.48
German <sup>d</sup>	95.88	83.10	96.27
Greek <sup>e</sup>	100.00	71.58	94.60
Italian <sup>f</sup>	100.00	86.05	95.77
Portuguese <sup>g</sup>	100.00	88.85	80.36
Spanish <sup>h</sup>	100.00	89.46	96.92

These numbers are gained from tests made by Explosion AI, the organization behind spaCy. They tested their language models with data similar to the data the models are based on, these are mostly Wikipedia text sources. This means that these numbers can not be assigned to all text genres we are processing. Until now we have not made accuracy tests with our data but we expect lower accuracies.

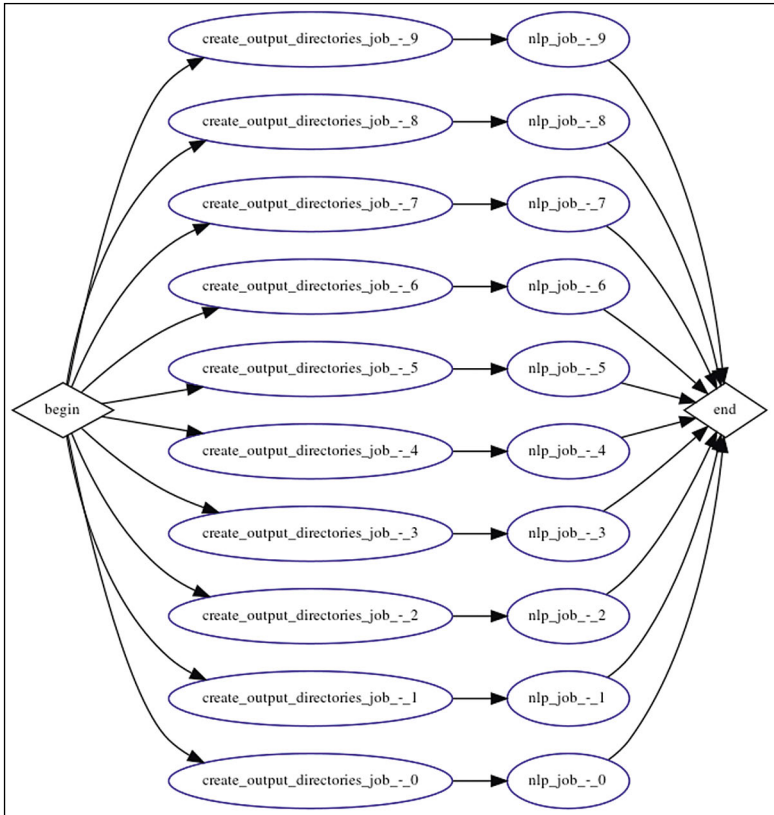
- a *Explosion AI*, Nl\_core\_news\_sm, version 2.1.0 (2019), [https://github.com/explosion/spacy-models/releases/tag/nl\\_core\\_news\\_sm-2.1.0](https://github.com/explosion/spacy-models/releases/tag/nl_core_news_sm-2.1.0).
- b *Explosion AI*, En\_core\_web\_sm, version 2.1.0 (2019), [https://github.com/explosion/spacy-models/releases/tag/en\\_core\\_web\\_sm-2.1.0](https://github.com/explosion/spacy-models/releases/tag/en_core_web_sm-2.1.0).
- c *Explosion AI*, Fr\_core\_news\_sm, version 2.1.0 (2019), [https://github.com/explosion/spacy-models/releases/tag/fr\\_core\\_news\\_sm-2.1.0](https://github.com/explosion/spacy-models/releases/tag/fr_core_news_sm-2.1.0).
- d *Explosion AI*, De\_core\_news\_sm, version 2.1.0 (2019), [https://github.com/explosion/spacy-models/releases/tag/de\\_core\\_news\\_sm-2.1.0](https://github.com/explosion/spacy-models/releases/tag/de_core_news_sm-2.1.0).
- e *Explosion AI*, El\_core\_news\_sm, version 2.1.0 (2019), [https://github.com/explosion/spacy-models/releases/tag/el\\_core\\_news\\_sm-2.1.0](https://github.com/explosion/spacy-models/releases/tag/el_core_news_sm-2.1.0).
- f *Explosion AI*, It\_core\_news\_sm, version 2.1.0 (2019), [https://github.com/explosion/spacy-models/releases/tag/it\\_core\\_news\\_sm-2.1.0](https://github.com/explosion/spacy-models/releases/tag/it_core_news_sm-2.1.0).
- g *Explosion AI*, Pt\_core\_news\_sm, version 2.1.0 (2019), [https://github.com/explosion/spacy-models/releases/tag/pt\\_core\\_news\\_sm-2.1.0](https://github.com/explosion/spacy-models/releases/tag/pt_core_news_sm-2.1.0).
- h *Explosion AI*, Es\_core\_news\_sm, version 2.1.0 (2019), [https://github.com/explosion/spacy-models/releases/tag/es\\_core\\_news\\_sm-2.1.0](https://github.com/explosion/spacy-models/releases/tag/es_core_news_sm-2.1.0).

## Implementation and workflow

Like our OCR implementation the NLP process is also implemented as a software pipeline. It is capable of processing text corpora in Dutch, English, French, German, Greek, Italian, Portuguese, and Spanish. It will only accept raw text files as input and provides verticalized text files as a result. The pipe-

line contains only one processing step, which is the spaCy natural language processing. While implementing this as a pipeline may sound laborious, it gives us the flexibility to easily extend the pipeline in the future.

Fig. 8: NLP pipeline procedure of ten input files



In order to achieve good computational performance, we aimed to make use of modern multicore systems. For that purpose, we used pyFlow, a powerful parallel task processing engine. Figure 8 shows one pipeline run where ten input files are processed. Each of these files is treated in a separate task that can run parallel to the others if the hardware is capable of doing parallel computation.

The NLP pipeline is deployed in a *Linux* container; for that purpose, we created a Dockerfile which tells the Docker build system to install all dependencies which are needed for using our NLP pipeline in a container image. The source code of this is available at the GitLab system hosted by Bielefeld University.<sup>77</sup> There you will find instructions on how to build and use the image or, in case you do not want to build the image yourself, we also offer an image in a prebuilt state.<sup>78</sup> The published image contains spaCy<sup>79</sup> in version 2.1.0 and language packages for processing Dutch, English, French, German, Greek, Italian, Portuguese and Spanish texts. All Software that is needed to realize and use this image is completely free and open source.

After the image was created, we were able to start multiple instances of the natural language processing software, encapsulated in Linux containers. Each instance executes one NLP pipeline run which processes the input data. An execution is bound to one specific text language, so the files processed within one pipeline run must contain texts in the same language.

Our usual workflow is described in the following:

1. Receive text corpora as raw text files
2. Create input and output directories
3. Copy files into the input directory
4. Start the NLP software
  - `nlp -i <inputdir> -l <languagecode> -o <outputdir>`
5. Check the results in the output directory

The results are saved as verticalized text files. This is a XML compliant format, where each line contains one token with all its assigned attributes. One line is structured in the following order: word, lemmatized word, simplified part-of-speech tag, part-of-speech tag, named entity recognition tag. NULL indicates that no named entity is recognized. The beginning and end of a sentence is represented by `<s>` and `</s>`

---

77 Jentsch, Patrick/Porada, Stephan, Natural Language Processing, version 1.0, 2019, [https://gitlab.uni-bielefeld.de/sfb1288inf/nlp/tree/from\\_text\\_to\\_data](https://gitlab.uni-bielefeld.de/sfb1288inf/nlp/tree/from_text_to_data) [accessed: 31.08.2019].

78 P. Jentsch/S. Porada, Docker Image: Natural Language Processing.

79 Explosion AI, SpaCy, version 2.1.0 (2019), <https://github.com/explosion/spaCy/releases/tag/v2.1.0> [accessed: 31.08.2019].



and analog to this you have a start and an end tag for the text and the complete corpus.

As a short example, we process the text “Tesseract is a software maintained by Google.” The resulting verticalized text file looks as follows:

```
<?xml version="1.0" encoding="UTF-8"?>
<corpus>
<text>
<s>
  Tesseract tesseract NOUN NN NULL
  is be VERB VBZ NULL
  a a DET DT NULL
  software software NOUN NN NULL
  maintained maintain VERB VBN NULL
  by by ADP IN NULL
  Google google PROPN NNP ORG
  . . PUNCT . NULL
</s>
</text>
</corpus>
```

## spaCy and Stanford CoreNLP

Before we decided to use spaCy as our natural language processing software we worked with Stanford CoreNLP. This software is versatile but does not offer its full functionality for all languages used in our context. In order to work with the processed texts, we decided to use a software called The IMS Open Corpus Workbench (CWB), which requires the verticalized text file format as input. With Stanford CoreNLP we had to write complex conversion programs to transfer the output into the desired verticalized text format in order to be able to import it to our CWB instance.

Table 3 shows our functional requirements and those supported by *Stanford CoreNLP* for specific languages.<sup>80</sup>

---

<sup>80</sup> *Stanford University*, Using Stanford Corenlp on Other Human Languages, <https://stanfordnlp.github.io/CoreNLP/human-languages.html#models-for-other-languages> [accessed: 27.03.2019].

Table 3: Functionality of Stanford CoreNLP

	English	French	German	Spanish
Tokenization	x	x		x
Lemmatization	x			
Part-of-Speech Tagging	x	x	x	x
Named Entity Recognition	x		x	x

To handle the mentioned problems we switched to spaCy, which offers the needed functionalities for all languages we encounter and also has a good application programming interface, which we were able to utilize in order to create the needed verticalized text files.

## 6. Conclusion

With our pipeline we hope to encourage scientific researchers, not only at Bielefeld University, to make use of the means provided by the digital age. Since the cooperation between information science and the humanities lies at the very core of Digital Humanities, we want to contribute our share by providing a tool to digitize and process texts. This tool is constantly revised and improved in accordance with the needs of its potential users and in close collaboration with them.

## Bibliography

- Adobe Systems Incorporated*, Document Management – Portable Document Format – Part 1: PDF 1.7 (2008), 25 ff., [https://www.adobe.com/content/dam/acom/en/devnet/pdf/PDF32000\\_2008.pdf](https://www.adobe.com/content/dam/acom/en/devnet/pdf/PDF32000_2008.pdf) [accessed: 26.03.2019].
- Adobe Systems Incorporated*, TIFF: Revision 6.0, version 6.0 (1992), 57–58, <https://www.adobe.io/content/dam/udp/en/open/standards/tiff/TIFF6.pdf> [accessed: 31.08.2019].
- Berlin-Brandenburgische Akademie der Wissenschaften*, Ziel und Fokus des DTA-Basisformats (Deutsches Textarchiv, Zentrum Sprache der Berlin-Brandenburgischen Akademie der Wissenschaften), <http://www.deutsches>

- textarchiv.de/doku/basisformat/ziel.html#topic\_ntb\_ssd\_qs\_\_rec [accessed: 12.03.2019].
- Boettiger, Carl, An Introduction to Docker for Reproducible Research, in: ACM SIGOPS Operating Systems Review 49 (2015), 71–79, <https://doi.org/10.1145/2723872.2723882>.
- Carrasco, Rafael C., Text Digitisation, <https://sites.google.com/site/textdigitisation/> [accessed: 11.06.2019]
- Chaudhuri, Arindam et al., Optical Character Recognition Systems for Different Languages with Soft Computing, Springer International Publishing: 2017, 90–92, 17–22, <https://doi.org/10.1007/978-3-319-50252-6>.
- Cheriet, Mohamed, et al., Character Recognition Systems, John Wiley & Sons, Inc.: 2007, 8–15, <https://doi.org/10.1002/9780470176535>.
- Debian, version 9 (The Debian Project, 2017), <https://www.debian.org/> [accessed: 31.08.2019].
- Docker, version 18.09.1 (Docker, 2013), <https://www.docker.com/> [accessed: 31.08.2019].
- Explosion AI, Linguistic Features · spaCy Usage Documentation: Tokenization, <https://spacy.io/usage/linguistic-features#tokenization> [accessed: 25.03.2019].
- Explosion AI, Industrial-Strength Natural Language Processing (NLP) with Python and Cython: SpaCy/Spacy/Lang/de/Lemmatizer.py, <https://github.com/explosion/spaCy/tree/v2.1.0/spacy/lang/de/lemmatizer.py> [accessed: 21.05.2019].
- Explosion AI, Annotation Specifications · spaCy API Documentation: Part-of-Speech Tagging, <https://spacy.io/api/annotation#pos-tagging> [accessed: 27.03.2019].
- Explosion AI, SpaCy, version 2.1.0 (2019), <https://github.com/explosion/spaCy/releases/tag/v2.1.0> [accessed: 31.08.2019].
- Google Inc., Tesseract OCR (2019), <https://github.com/tesseract-ocr/tesseract/> [accessed: 31.08.2019].
- Google Inc., ImproveQuality: Improving the Quality of the Output (2019), <https://github.com/tesseract-ocr/tesseract/wiki/ImproveQuality> [accessed: 31.08.2019].
- Hardie, Andrew/Evert, Stefan, IMS Open Corpus Workbench, <http://cwb.sourceforge.net/> [accessed: 31.08.2019].
- International Organization for Standardization, ISO/Iec 9126-1:2001: Software Engineering – Product Quality – Part 1: Quality Model (International

- Organization for Standardization, June 2001), <https://www.iso.org/standard/22749.html> [accessed: 31.08.2019].
- International Organization for Standardization*, ISO/Iec 25010:2011: Systems and Software Engineering – Systems and Software Quality Requirements and Evaluation (SQuaRE) – System and Software Quality Models (International Organization for Standardization, March 2011), <https://www.iso.org/standard/35733.html> [accessed: 31.08.2019].
- Jackson, Mike/Crouch, Steve/Baxter, Rob*, Software Evaluation: Criteria-Based Assessment (Software Sustainability Institute, November 2011), <https://software.ac.uk/sites/default/files/SSI-SoftwareEvaluationCriteria.pdf> [accessed: 31.08.2019].
- Jentsch, Patrick/Porada, Stephan*, Docker Image: Optical Character Recognition, version 1.0, Bielefeld 2019, [https://gitlab.ub.uni-bielefeld.de/sfb1288inf/ocr/container\\_registry](https://gitlab.ub.uni-bielefeld.de/sfb1288inf/ocr/container_registry) [accessed: 31.08.2019].
- Jentsch, Patrick/Porada, Stephan*, Natural Language Processing, version 1.0, 2019, [https://gitlab.ub.uni-bielefeld.de/sfb1288inf/nlp/tree/from\\_text\\_to\\_data](https://gitlab.ub.uni-bielefeld.de/sfb1288inf/nlp/tree/from_text_to_data) [accessed: 31.08.2019].
- Jurafsky, Daniel/Martin, James H.*, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Draft of September 23, 2018, 2018, 151, 156.
- Manning, Christopher D./Raghavan, Prabhakar/Schutze, Hinrich*, Introduction to Information Retrieval, Cambridge: Cambridge University Press, 2008, 22, <https://doi.org/10.1017/cbo9780511809071>.
- Olah, Christopher*, “Understanding LSTM Networks, <https://colah.github.io/posts/2015-08-Understanding-LSTMs/> [accessed: 26.03.2019].
- Oztan, Bazak, et al.*, Removal of Artifacts from JPEG Compressed Document Images, in: Reiner Eschbach/Gabriel G. Marcu(eds.), Color Imaging XII: Processing, Hardcopy, and Applications, (SPIE) 2007, 1-3, <https://doi.org/10.1117/12.705414>.
- Red Hat Inc.*, What’s a Linux Container?, <https://www.redhat.com/en/topics/containers/whats-a-linux-container> [accessed: 20.05.2019].
- Saunders, Chris*, PyFlow: A Lightweight Parallel Task Engine, version 1.1.20, 2018, <https://github.com/Illumina/pyflow/releases/tag/v1.1.20> [accessed: 31.08.2019].
- Smith, Ray*, Building a Multilingual OCR Engine: Training LSTM Networks on 100 Languages and Test Results (Google Inc., June 20, 2016), 16, 17, <https://>

- [github.com/tesseract-ocr/docs/blob/master/das\\_tutorial2016/7Building%20a%20Multi-Lingual%20OCR%20Engine.pdf](https://github.com/tesseract-ocr/docs/blob/master/das_tutorial2016/7Building%20a%20Multi-Lingual%20OCR%20Engine.pdf).
- Smith, Ray*, An Overview of the Tesseract OCR Engine, in: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2 (2007), <https://doi.org/10.1109/icdar.2007.4376991>.
- Smith, Ray*, A Simple and Efficient Skew Detection Algorithm via Text Row Accumulation, in: Proceedings of 3rd International Conference on Document Analysis and Recognition (IEEE Comput. Soc. Press, 1995), <https://doi.org/10.1109/icdar.1995.602124>.
- Sourceforge*, CWB/Perl & Other APIs, [http://cwb.sourceforge.net/doc\\_perl.php](http://cwb.sourceforge.net/doc_perl.php) [accessed: 13.05.2019].
- Stanford University*, Using Stanford CoreNlp on Other Human Languages, <https://stanfordnlp.github.io/CoreNLP/human-languages.html#models-for-other-languages> [accessed: 27.03.2019].
- Text Encoding Initiative*, TEI P5: Guidelines for Electronic Text Encoding and Interchange (TEI Consortium) <https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf> [11.06.2019].
- University of Alicante*, ocrevalUAtion, <https://github.com/impactcentre/ocrevalUAtion> [accessed: 05.04.2019].
- University of Alicante*, OcrevalUAtion, version 1.3.4 (2018), <https://bintray.com/impactocr/maven/ocrevalUAtion> [accessed: 31.08.2019].
- Ye, Peng/Doermann, David*, Document Image Quality Assessment: A Brief Survey, in: 2013 12th International Conference on Document Analysis and Recognition (2013), 723, <https://doi.org/10.1109/icdar.2013.148>.

### **III. Digital Research Perspectives from Different Humanities Disciplines**



# Testing Hypotheses with Dirty OCR and Web-Based Tools in Periodical Studies<sup>1</sup>

---

Malte Lorenzen

Periodical studies always played only a marginal role in the philologies. But for a couple of years there has been at least a small boom in research on journals and magazines which depends to a large extent on the emergence of the digital humanities.<sup>2</sup> The opportunity to browse through large corpora and search with ease for the rise and fall of distinct topics, or even for the use of single words over time made it once again attractive to deal with a type of text which was formerly often rejected as object for research because of its various contents, its ephemerality and its apparent distance from high literature.

The necessary condition of any research with the means of the digital humanities is the availability of a digital corpus that can be processed by computer tools. Yet this availability is still one of the most urgent problems. Though much work in the process of digitizing hundreds of years of print culture has been done, there is no guarantee that the very texts one needs for a project are available anywhere in the World Wide Web.<sup>3</sup> And even if

---

1 I am very grateful to Kai Kauffmann for our discussions on the possibilities and restrictions of tools and to Christine Peters and Joris C. Heyder for their critical readings of this article. Special thanks go to Laura Säumenicht for the digitization of the examined corpus.

2 A more or less initializing text for the new interest in periodicals as an object of the (digital) philologies is: *Latham, Sean, Scholes, Robert*, The Rise of Periodical Studies, in: *PMLA* 121 (2006), 517–531.

3 Cf. *Hahn, Carolin*, Forschung benötigt Infrastrukturen: Gegenwärtige Herausforderungen literaturwissenschaftlicher Netzwerkanalysen, in: Toni Bernhart et al. (ed.), *Quantitative Ansätze in den Literatur- und Geisteswissenschaften: Systematische und historische Perspektiven*, Berlin and Boston: de Gruyter, 2018, 315–34, esp. 326–28.



someone has already digitized the corpus in question, some of the texts might be nothing more than “a garbled mess” because of bad optical character recognition (OCR)<sup>4</sup> – the best known technique to bring printed texts into a machine readable form – and/or the condition of the underlying historical documents.<sup>5</sup> Not least because digitization is actually a very time consuming and, therefore, expensive process, some of the recent discussions in the digital humanities aim not only at ways to improve OCR but also at the question “how accurate [...] digitized sources [must] be to produce robust results”.<sup>6</sup> In other words, the discussion is about the possibility to obtain new scientific insights while working with ‘dirty’ OCR full of errors.

Previous studies were quite optimistic in this respect.<sup>7</sup> So as I am bound to a quite large corpus of documents in a research project on comparisons of cultures in German periodicals during World War I, I decided on digitizing at least a subcorpus to see how far I can get with it despite possible OCR errors. The main interest of this article is to map out the potential of provisional digital documents. Thus, this contribution has a somewhat experimental character because it not only rests upon dirty OCR but also upon capabilities of tools.

Apart from the availability of digital documents, the decision on software poses the most urgent problem for any digital philologist – even more so for any ‘newbie’ in Digital Humanities. When there is no research community including specialists in computer sciences who can develop software exactly fitting the interests in a given corpus of documents, it is the best option to choose free available tools and toolkits like *Voyant Tools*<sup>8</sup> or *AntConc*<sup>9</sup> with a

---

4 Cf. *Nicholson, Bob*, Counting Culture; or, How to Read Victorian Newspapers from a Distance, in: *Journal of Victorian Culture* 17 (2012), 242.

5 Cf. *Cordell, Ryan*, ‘Q i-jtb the Raven’: Taking Dirty OCR Seriously, in: *Book History* 20(2017), 194–95 and *Holley, Rose*, How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs, in: *D-Lib Magazine* 15 (2009), <https://doi.org/10.1045/march2009-holley>.

6 *Strange, Carolyn et al.*, Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers, in: *Digital Humanities Quarterly* 8 (2014), <http://www.digitalhumanities.org/dhq/vol/8/1/000168/000168.html>.

7 Cf. *ibid.*

8 <https://voyant-tools.org/> [accessed: 14.05.2019].

9 <http://www.laurenceanthony.net/software/antconcl/> [accessed: 14.05.2019].

quite easily understandable interface.<sup>10</sup> This, in turn, has the deficiency that there is no opportunity to modify the software in relation to one's questions – at least not for anyone with a lack of suitable software skills. So quite the contrary becomes necessary, one's own questions have to be fitted to the software.<sup>11</sup>

In the following essay, I will (1) present the process of creating a digital corpus and possibilities to test the functionality of dirty OCR. Then (2), two different approaches of working with digital tools will be shown. First (2.1), data mining as a process of searching for trends and patterns without any strong presuppositions will be introduced. Second (2.2), the ability of dirty OCR and digital tools will be checked when (re)examining hypotheses gained by close or surface reading.<sup>12</sup> Essential questions will focus on the validity of produced data and what kind of research issues can be handled with dirty OCR and free available tools. In the end (3), there will be an answer to the question if it is worth investing time and work into digitization when the outcome is unavoidably provisional and erroneous.

## 1. The creation of a digital corpus

Pretty soon after deciding to give Digital Humanities a try it became clear to me that it would be impossible to digitize the whole corpus of four to five volumes of six different periodicals – comprising at least 30,000 pages.<sup>13</sup> So, a selection had to be made. Because much of the conception of the project is based on close and surface reading of *Süddeutsche Monatshefte*,<sup>14</sup> said peri-

---

10 For a further discussion of open tools cf. the contribution of Helene Schlicht in this volume.

11 For a discussion of the “epistemological proposition” of any tool, cf. *Rieder, Bernhard/Röhle, Theo*, Digital Methods: Five Challenges, in: David M. Berry (ed.), *Understanding Digital Humanities*, New York: Palgrave Macmillan, 2012, 67–84, esp. 68–71.

12 Cf. for a plea for “surface reading” as a third mode of reading between “close” and “distant reading” *Collier, Patrick*, What is Modern Periodical Studies?, in: *Journal of Modern Periodical Studies* 6 (2015), 107–108.

13 The whole print corpus consists of the war year volumes of *Neue Rundschau*, *Deutsche Rundschau*, *Die Zukunft*, *Die Gegenwart*, *Die Gesellschaft*, and *Süddeutsche Monatshefte*.

14 For an overview of some of the main hypotheses of the project cf. *Kauffmann, Kai*, *Wissensvermittlung, Kulturpolitik und Kriegspropaganda: Thesen zur Kriegspublizistik der*

odical was chosen. The plan was to not only validate the results of the digital tools through a comparison with the results of a human reader but also, vice versa, to check the results and main hypotheses of the human reader against those of the reading machine. Despite this limitation, there still remained around 7,000 pages to be digitized.

All these pages had to be scanned manually. Everyone who ever made scans for reading on screen knows that it is not that bad when pages are slightly crooked but it might lead to serious problems when those scans are meant to be further processed for the means of digital tools. The issues of *Süddeutsche Monatshefte* that were accessible are bound book-like volumes with up to 1,000 pages each. Especially in the middle of such big books usual scanners struggle with the book fold which can produce slanted or blurry pictures. So we not only needed to scan every single page but also had to review every digitized page, and, if necessary, scan it again or straighten it digitally, which required additional hours of work.

The scanned PDF-files were then transformed with OCR, using *Tesseract*, into raw text files with no further mark-ups and annotations.<sup>15</sup> Figures 1a and 1b show an original scan alongside the OCR transformed .txt-version that served as a basis for the work with digital tools.<sup>16</sup> Obviously, this is far off the demands of editorial philology and would be wholly inadequate as a basis for a digital or non-digital edition in the proper sense.<sup>17</sup> To make the results suitable for that, it would actually be necessary to invest even more time to correct the errors, which would still have to be conducted manually for the most part, despite all progress in training software to reach better results. But as time and resources were limited, that could not be done in this project.

---

deutschen Rundschauzeitschriften 1914–1918, in: Olivier Agard/Barbara Beßlich (eds.), *Krieg für die Kultur? Une guerre pour la civilisation? Intellektuelle Legitimationsversuche des Ersten Weltkriegs in Deutschland und Frankreich (1914–1918)*, Berlin: Peter Lang, 2018, 113–128.

- 15 For an expert description of this process, see the article of Patrick Jentsch and Stephan Porada in this volume.
- 16 For methods of measuring OCR accuracy that were not applied here cf. *R. Holley, How Good Can It Get?*
- 17 At the present moment, the standard for digital editions is defined by the guidelines of the Text Encoding Initiative (TEI); cf. <https://tei-c.org/guidelines/> [accessed: 14.05.2019].

Fig. 1a: Original scan of a magazine page

## Die Geschichte der Ostseeprovinzen.

Von Theodor Schiemann, Professor der Geschichte an der Universität Berlin.

Durch die Geschichte Osteuropas zieht vom 13. Jahrhundert bis in die Gegenwart ein noch ungelöstes Problem: das Ringen um die Beherrschung der Ostsee, die Frage des *Dominium maris Baltici*. Die Herrschaft auf dem Baltischen Meer, der Ostsee, und damit die Vorherrschaft in Osteuropa, gehört demjenigen, der die Küsten sich zu eigen macht, oder, historischer formuliert, dem Herrn der heutigen Ostseeprovinzen Rußlands, die man noch bis in die Gegenwart hinein die „Deutschen Ostseeprovinzen Rußlands“ nennt. Die heute so geläufigen Bezeichnungen Valte und Baltische Provinzen sind erst nach 1860 aufgekomen. Bis dahin schrieb und sprach man wohl vom Baltischen Meer, aber der Name wurde nicht auf die Küsten übertragen und ebensowenig auf die Bewohner des Landes. Die älteste Bezeichnung des Landes war Livonia, Livland, so genannt nach dem finnischen Stamm der heute fast ausgestorbenen Liven an der Küste nördlich der Düna. Verwandt waren ihnen die südlich am Meeresufer wohnenden Kuren, deren Name im heutigen Kurland fortlebt, und die Esten, die im heutigen Estland und im nördlichen Livland ihre Sitze hatten. Diese finnischen Stämme hatten auch die Inseln vor dem Rigaschen Meerbusen inne und waren gefürchtete Seeräuber. Sie brandschatzten die skandinavischen Küsten, ganz wie die slawischen Stämme im heutigen Pommern und Mecklenburg den westlichen Teil der Ostsee für die Seefahrt gefährdeten. Erst die deutschen Orlogschiffe haben dort allmählich einen Seefrieden herzustellen vermocht.

Nördlich von den finnischen Stämmen der Küste lagen die Sitze der den Preußen und Litauern nahe verwandten Letten, die von ihren kriegs- und heutelustigen Nachbarn arg bedrängt wurden. Dank dem Schutz der Deutschen sind ihnen später allmählich die Sitze der zusammenschmelzenden Stämme der Liven und Kuren zuteil geworden, während die zäheren Esten sich nicht nur auf ihrem ursprünglichen Boden behauptet, sondern weiter nach Süden ausgebreitet haben.

Im Rücken all dieser größeren und kleineren Volksplitter saßen Russen, die Fürsten von Pologk, deren Einflußsphäre bis kurz vor Riga reichte, weiter nördlich die beiden Stadtrepubliken Pskow und Nowgorod. Ihnen, so schien es, mußte die Herrschaft über die minder wehrhaften, noch heidnischen Bewohner der Ostseeküste und damit die Anwartschaft auf das künftige *Dominium maris Baltici* zufallen. So war die Lage um die Mitte des 12. Jahrhunderts.

Da haben die Deutschen eingegriffen, und zwar die drei lebendigsten Faktoren des deutschen Mittelalters: das städtische Bürgertum, die

Fig. 1b: The OCR transformed .txt-version of the same page

597

Die Geschichte der Ostseeprovinzen

Von Theodor Schiemann, Professor der Geschichte an der Universität Berlin.

Durch die Geschichte Osteuropas zieht vom 13. Jahrhundert bis in die Gegenwart ein noch ungelöstes Problem: das Ningen um die Beherrschung der Ostsee, die Frage des Dominium maris Baltici. Die Herrschaft aus dem Baltischen Meer, der Ostsee, und damit die Borherrschaft in Osteuropa, gehört demjenigen, der die Küsten sich zu eigen macht, oder, historischer formuliert, dem Herrn der heutigen Ostseeprovinzen Nußlands, die man noch bis in die Gegenwart hinein die „Deutschen Ostseeprovinzen Rußlands« nennt. Die heute so geläufigen Bezeichnungen Balte und Bal-tische Provinzen sind erst nach 1860 aufgekommen. Bis dahin schrieb und sprach man wohl vom Baltischen Meer, aber der Name wurde nicht auf

die Küsten übertragen und ebensowenig auf die Bewohner des Landes.

Die älteste Bezeichnung des Landes war Livonia, Livland, so genannt nach dem finnischen Stamm der heute fast ausgestorbenen Liven an der -Küste nördlich der Düna. Verwandt waren ihnen die südlich am Meeres-ufer wohnenden Kuren, deren Name im heutigen Kurland fortlebt, und die Esten, die im heutigen Estland und im nördlichen Livland ihre Sitze hatten. Diese finnischen Stämme hatten auch die Jnseln vor dem Nigaschen Meerbusen inne und waren gesürchtete Seeräuber: Sie brandschatzten die skandinavischen Küsten, ganz wie die slawischen Stämme im heutigen Pommern und Mecklenburg den westlichen Teil der Ostsee für die See-fahrt gefährdeten. Erst die deutschen Örologsschiffe haben dort allmählich einen Seesrieden herzustellen vermocht-

Ostlich von den finnischen Stämmen der Küste lagen die Sitze der den Preußen und Litauern nahe verwandten Letten, die von ihren kriegs- und beutelustigen Nachbarn arg bedrängt wurden. Dank dem Schutz der Deut-schen sind ihnen später allmählich die Sitze der zusammenschmelzenden Stämme der Liven und Kuren zuteil geworden, während die zäheren Esten sich nicht nur aus ihrem ursprünglichen Boden behauptet, sondern weiter nach Süden ausgebreitet haben.

Jm Rücken all dieser größeren und kleineren Volkssplitter saßen Russen, die Fürsten von Polozk, deren Einflußsphäre bis kurz vor Nigareichte, weiter nördlich die beiden Stadtrepubliken Pskow und Nowgorod. Ihnen, so schien es, mußte die Herrschaft über die minder wehrhaften, noch heidnischen Bewoh-ner der Ostseeküste und damit die Anwartschaft aus das künftige Dominium maris Baltici zufallen. So war die Lage um die Mitte des 12. Jahrhunderts:

Oa haben die Deutschen eingegriffen, und zwar die drei lebendigsten Faktoren des deutschen Mittelalters: das städtische Bürgertum, die

Fortunately, many of the OCR-induced errors are recurring and can thus be taken into account in the use of text mining tools. Among the most common errors – typically in the OCR of Fraktur fonts – are the transformation of the capital “R” to a capital “N”, the capital “I” to a capital “J” and the lower case “s” to a lower case “f” or vice versa – though these transformation errors do not occur anytime.<sup>18</sup> So if you want to find results including the term “Rußland”

<sup>18</sup> Typical errors further involve umlauts and end-of-line hyphenation; cf. figure 1b with the highlighting of some typical errors; cf. also Riddell, Allen Beye, How to Read 22, 198 Journal

(Russia), for example, it is no problem to search for “Rußland” or the incorrect “Nußland”.<sup>19</sup>

Since text mining tools are not only valuable because they provide the possibility to search for single terms but also because they can identify structures, patterns and trends not yet recognized, there was need for closer scrutiny of the potential of the dirty OCR files. I worked with Voyant Tools for the most part, a web-based open-source software for computer philologist text analysis that offers a variety of tools and visualization.<sup>20</sup> As a basis for testing, the whole corpus was split up into fifty files representing the fifty scanned issues of *Süddeutsche Monatshefte* to find out if Voyant was able to identify the regional focus of them.<sup>21</sup> It was first browsed and then searched for occurrences of “Schweiz” (Switzerland). The results were striking and revealed a peak for the May 1916 issue, which mainly focuses on “Die Schweiz im Krieg” (“Switzerland at War”) and is the only issue with a focus on Switzerland in the corpus. Similar results with similar preconditions were reached when browsing for “Vatikan” (Vatican) and “Spanien” (Spain). When searching for “England” and “Frankreich” (France), two of Germany’s main enemies during World War I, things got a bit more blurry due to the significant rise of results. But comparing the tools results with the printed tables

---

Articles: Studying the History of German Studies with Topic Models, in: Lynne Tatlock/Matt Erlin (eds.), *Distant Readings: Topologies of German Culture in the long nineteenth Century*, New York: Rochester, 2014, 95.

- 19 It proved to be irrelevant whether both words were searched separately or together in form of a regular expression.
- 20 See footnote 7. The decision fell for Voyant instead of AntConc – which provide the same tools to a large extent –, on the one hand because of its more extensive visualization capabilities and, on the other hand and particularly, because of its integrated stop word list that greatly facilitates its use.
- 21 Since the beginning of the war, every issue of *Süddeutsche Monatshefte* had a main topic, ranging from domestic affairs to economic problems or geographical regions. For a schematic table of contents cf. K. Kauffmann, *Wissensvermittlung, Kulturpolitik und Kriegspropaganda*, 121–22. There would have been other ways of testing the functionality of the dirty OCR, of course, including automatized techniques. Anyway, the latter would have required a ‘clean’ subcorpus for comparison. So the decision fell for checking it manually. Searching for country names turned out to be a good way, firstly, because of the focus of the research project on comparisons between cultures based on nation states and, secondly, because of the availability of the tables of contents which often reveal the regional focus in their header.

of contents showed that Voyant was still able to identify those issues in the corpus with a focus on England and France.

Now how about the search for occurrences of “Russia” with the above-mentioned complications in the transformation into OCR readable text? Figure 2 is a visualization of the raw frequencies of “Rußland” and “Nußland” in the whole corpus. It shows significant peaks for the February, July and December 1915, March 1916 and January 1917 issues. Using the tables of contents as a basis for testing again, the results are mixed. On the one hand, Voyant identifies some of the special issues on Russia<sup>22</sup> and Russia as a thematic priority in other issues;<sup>23</sup> on the other hand, Voyant shows the highest peak for occurrences of “Rußland” and “Nußland” for the January 1917 issue. Though the guiding theme of this issue is “Äußere Politik” (foreign policy), judging by their titles, none of the articles seems to focus on Russia. A closer look at the text with the guiding help of the reader function of Voyant and the visualization of the distribution of “Rußland” and “Nußland” in the text can reveal the reason for this surprising insight. It is owed to an article by Graf Ernst Reventlow on the Turkish straits and their development.<sup>24</sup> Of course, anyone with enough knowledge about the Ottoman Empire or Turkey in the 19th century and during World War I could suppose the importance of Russia in such an article when finding it in the table of contents; nevertheless, this example shows the potential of digital tools to reveal what might remain hidden to a cursory look.

The most irritating outcome was undoubtedly the result for the October 1918 issue on the first anniversary of the Bolshevik Revolution in Russia. What was expected to be one of the highest peaks in the visualization is in fact only of medium height. At first, I was apt to blame it on dirty OCR. But that was wrong. Instead, the document length is responsible for the outcome. As Voyant shows in its summary section, the October 1918 issue belongs to the five shortest issues in the whole corpus – probably due to paper shortage

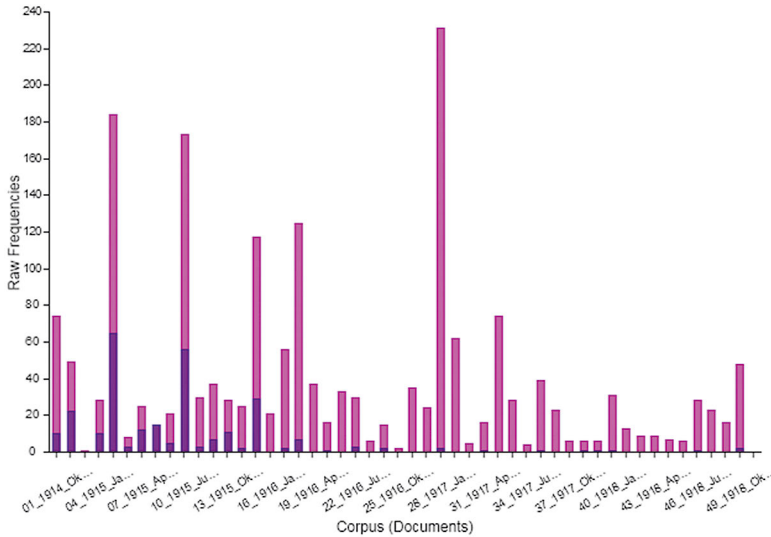
---

22 The February 1915 issue on “Rußland” (“Russia”) and the July 1915 issue on “Rußland von Innen” (“Russia from the Inside”).

23 This applies to the December 1915 issue, whose main topic are “Kriegsziele” (“War Aims”) and which contains at least some articles on Russia, and to the March 1916 issue, whose title is “Kriegsgefangen” (“War Captivity”) and which features many articles on Russia.

24 *Reventlow, Ernst, Die Frage der türkischen Meerengen und ihre Entwicklung*, in: *Süddeutsche Monatshefte* 14 (1917): 432–66.

Fig. 2: A visualization of the raw frequencies of “Rußland” and “Nußland” in the whole corpus



towards the end of the war – with an amount of 35,463 (recognized) tokens – in contrast, the longest issue has an amount of 90,518 (recognized) tokens.

So this and the previous example do not reveal that much about the quality of the OCR and its usability but about the formal structure and the contents of the periodical it is based on. First, there are the above-mentioned differences in the length of each document representing a single issue that produce results different from those we would expect with the tables of content in mind.<sup>25</sup> In this case, the described results are no huge surprise, for sure, but they could have occurred with other words and in other contexts.<sup>26</sup>

25 A possible solution to this problem might be the calculation of the median value; cf. *Jannidis, Fotis/Kohle, Hubertus/Rehbein, Malte (eds.)*, *Digital Humanities: Eine Einführung*, Stuttgart: Springer, 2017, 282–83. Nevertheless, the inexperience of most traditionally educated humanists with statistics and mathematics comes into play here. Anyone engaging deeper in Digital Humanities has to learn whole new things to check the data for its reliability.

26 See, for example, *Roff, Sandra*, *From the Field: A Case Study in Using Historical Periodical Databases to Revise Previous Research*, in: *American Periodicals 18* (2008), 96–100. In her brief account, she emphasizes the great opportunities of keyword searches in



Second, a single article within the issues of an outstanding length might lead to some distortions in the results. This is the case with the above-mentioned article on the Turkish straits and is more often the case in *Süddeutsche Monatshefte* and similar magazines.<sup>27</sup>

The OCR itself is certainly not good enough to identify each and every token of a given word. But considering the fact that Digital Humanities is used for the recognition of larger structures and patterns, too, this is not necessary at all. The dirty OCR seems capable of identifying such trends and might thus be suited to serve as an instrument at least for testing hypotheses developed through close readings of selected parts or surface readings of larger parts of the corpus in question.

## 2. Working with dirty OCR

Ultimately, all tools for text analysis in Digital Humanities serve two different purposes. On the one hand, they can search Big Data for patterns and structures which are not discovered yet or are not to be discovered at all with the means of traditional reading techniques – this is generally known as “data mining”. On the other hand, insights gained by traditional reading techniques can be tested with the means of Digital Humanities.<sup>28</sup> Though both approaches can be combined, it is obvious that both of them require different forms of engagement. While one can simply use the abilities of tools to reveal frequencies or patterns for further investigation in the first case, it is inevitable to consider beforehand what one actually wants to find out in the second case. Both options will be tested in the following sections for their hermeneutical significance as well as for their potential with a corpus full of OCR noise.

---

big databases, where previously only titles of articles served as hints for further investigation.

27 There is a significant difference between newspapers and book-like journals: The former are not as prone to distortions because of their manifold content and the smaller extent of their articles.

28 In a slightly different context this is the differentiation between a “corpus-driven approach” and a “corpus-based approach”; cf. *Anthony, Laurence*, A Critical look at software tools in corpus linguistics, in: *Linguistic Research* 30 (2013), 141–61.

## 2.1 Data mining with dirty OCR

The text corpus was uploaded to Voyant without any problems. One of the first things it provides is a word cloud which displays the most frequent words.<sup>29</sup> As figure 3 shows, dirty OCR definitely causes problems.

Fig. 3: A word cloud without additions to the stop list ...



This becomes most obvious when looking at single characters like “i”, “a” and “e” and at the German prefix “ge-”, probably split off from words because of bad character recognition or problems with end-of-line hyphenation. Further, there are words like “jn” and “jch”, correctly spelled with an “i”. The words “in” (in) and “ich” (I) are usually detected by the integrated stop list of Voyant, which, to a large extent, consists of function words with less hermeneutical significance. Due to errors in the character recognition, which transformed the Fracture “i” into a “j”, the stop list does not work correctly either. This is undoubtedly annoying but still a manageable problem. It is quite easy to adjust the stop list by adding single characters, prefixes and numbers. After making additions to the stop list for each error and five iter-

29 It is only the font size and position in the cloud that reveals the importance of the words due to their frequency; the different colors of the words produced by Voyant are of no significance and thus not depicted here.

ations, I arrived at the word cloud shown in figure 4. This cloud is free from OCR noise and could serve as a starting point for interpretation.<sup>30</sup>

Fig. 4: ... and with additions to the stop list



Apparently, there are some commonalities between both word clouds. On top, there is the central position and large point size of words with the stem “deutsch-” (German).<sup>31</sup> What does that mean? Well, this is the point where text mining stops and interpretation begins because no tool will ever tell anything about the hermeneutical significance of its outcome. In this case, the results might not be very surprising at first sight. Due to the fact that the

30 For a plea for digital corpus linguistics including especially the counting of word frequencies as a starting point for further investigations see Archer, Dawn, Data Mining and Word Frequency Analysis, in: Gabriele Griffin/Matt Hayler (eds.), *Research Methods for Reading Digital Data in the Digital Humanities*, Edinburgh: Edinburgh University Press, 2016, 72–92.

31 Another problem becomes obvious, here. It is the lack of any further lemmatization in the raw text files as well as in Voyant's stop list. This means, that any deviant grammatical form of a given word is recognized as a new type of word. To some extent, this problem could be solved with the use of regular expressions though this is quite time consuming as well.



In a project concerned with publicists comparing nations and cultures, the result can be interpreted as that the growing interest in other nations during World War I does not constitute an actual interest in the nations themselves. Instead, every statement and opinion piece about other countries contained in the journal seems to only be relevant in relation to Germany. The nature of these relationships is, however, not visible in the word cloud but needs to be further investigated by means of other tools<sup>33</sup> or through close readings of single articles and issues.<sup>34</sup>

There are some more commonalities between the word clouds in figure 4 and 5, in particular the visible significance of the main German war opponents England, France and Russia<sup>35</sup> – and the United States? It is surprising that they are not part of this cloud, even in view of the fact of their late entry into the war in 1917. At least there are two issues of *Süddeutsche Monatshefte* that have a focus on the USA and some more articles in other issues. Are they missing because of dirty OCR? I doubted that because in general the character recognition worked quite well; instead, at some point, I doubted the functionality of Voyant and presumed it stopped processing the corpus somewhere in the middle.<sup>36</sup> This assumption was disproved, however, when I finally found the reason, namely the lack of any further annotation, espe-

---

33 Some possibilities for similar problems will be discussed in section 2.2 below.

34 Though it is likely, that comparisons are quite important within the constructed and presented relations, the reasons for such a strong orientation towards similarities and distinctions across nations are not so evident. Two main aspects might be at work: On the hand, the situation of war might play a role. Measures and means of administrating occupied territories have to be discussed as well as strengths and weaknesses of the enemies to assess the possibility of victory or probable risks that could undermine victory. On the other hand, a more general aspect might be at work. Comparing other nations and/or cultures with one's own can serve as a means to arouse interest or to help understand the 'other'.

35 "England", "Frankreich", "Rußland".

36 In fact, Voyant might have problems like that in some cases due to server capacities and limited working memory – a problem that arose in my work with Voyant when trying out its tool for topic modeling. More general, this leads to the question of epistemic trust in the tools' functionality: As long as one is not able to read the codes of the tool, there remains nothing more than to trust in the produced data. However, there is a quite simple solution that might help at least in the case of large corpora where it is not possible to check the results with close reading: the use of "different tools from the same category" – cf. B. Rieder/T. Röhle, *Digital Methods*, 77.

cially Named Entity Recognition (NER) and lemmatization. A closer look at the word cloud reveals the inconspicuous word “staaten” (states). What could simply be the plural form of “Staat” (state) – and in some cases stands for nothing more – turns out to be a part of the term “vereinigte staaten” (united states) in many instances.<sup>37</sup> Without annotation, Voyant treats “Vereinigte Staaten” (United States) not as a single term but splits it into two terms. Moreover, when testing for instances of “amerika” (America), which in most cases is used synonymously with “United States”, the term turned out to appear only 701 times in the corpus. In comparison, “frankreich” (France) is used 1976 times. Then again, “amerika\*” can be found 2,074 times in the document. This clearly shows that digital tools without annotated corpora do have their limits when processing inflectional languages like German. It also demonstrates the need to pay close attention to the functioning of any tool and to the condition of any digitized corpus. This is especially true when one compares prior knowledge – or rather expectations – to actual results. Great differences between them do not necessarily rest upon wrong expectations but could be the result of a malfunctioning tool or somehow flawed documents.

Let us turn from here to the differences between figure 4 and 5. One of them is visible at the bottom left of figure 5. “belgien” (Belgium) is missing in the word cloud displaying all the issues and comes into play only in the word cloud displaying the issues on other countries than Germany. Moreover, it is the only smaller nation involved in World War I which is shown<sup>38</sup> – while even Germany’s most important ally, Austria, is missing. This intriguing result leaves room for speculation. Was Belgium more important to the authors of *Süddeutsche Monatshefte* than Austria because it was German-occupied and they therefore felt the need to discuss means of administration in the Belgian territory? Or was it because allied reports on German atrocities in Belgium had to be denied?<sup>39</sup>

---

37 This was tested with the help of Voyant’s contexts tool – but any tool with the possibility to show the amount of counted words would work as well – with the result of 426 instances of “vereinigte staaten” and “vereinigten staaten” and a total number of 1,150 instances of “staaten”.

38 The word would have been depicted even bigger if Voyant had integrated instances of “velgien”, which stems from bad OCR.

39 For the international discussion of German war atrocities in Belgium cf. the extensive study of *Horne, John/Kramer, Alan: German Atrocities, 1914: A History of Denial*, New Haven: Yale University Press, 2002.

In fact, the result is misleading, at least partially. Using Voyant’s document terms tool to get an overview of word occurrences reveals that instances of “\*sterreich”<sup>40</sup> (836 instances) exceed instances of “\*elgium” (699) in the whole text corpus. Only in the subcorpus containing issues on other nations than Germany does the use of “\*elgium” (631) exceed instances of “\*sterreich” (529). It can be concluded that overall Belgium is not considered more important than Austria. Instead, this comparison reveals a tendency to write articles on Austria not so much in contexts of foreign nations but in relation to the writers’ own German nation.

Nevertheless, the speculations about the importance of Belgium to German writers still stand. The depicted word cloud provides no indication of what the articles are about, however. To shed light on this matter, I cut every article on Belgium out of the whole text corpus,<sup>41</sup> uploaded the collection to Voyant and arrived at the result displayed in figure 6.

Fig. 6: Word cloud of any article on Belgium



- 40 Searching for “österreich” (Austria) turned out to be not the best option because of the initial umlaut that leads to OCR errors in many instances.
- 41 This step was necessary partly because issues on Belgium had more central topics – for example, the April 1915 issue with another focus on Bismarck at the occasion of his hundredth birthday –, partly because articles from other issues could be taken into account. The selection was done by the guidance of the table of contents so that any article with a mention of Belgium in its title was included in this subcorpus.

What strikes the attention in this word cloud are – besides the expectable central position and font size of words with the stem “belg\*”, once again, the high frequency of words with the stem “deutsch\*” (German), the occurrence of France and England as Germany’s main enemies at its western front, and the terms “krieg” (war) and “neutralität” (neutrality) – essentially two things. The first is a concentration of urban spaces like “brüssel” (Brussels), “antwerpen” (Antwerp) and “küste” (coast). While this clearly stems from the requirements and conditions of the war, there is furthermore a focus on the main ethnic groups in Belgium, namely “flamen” or “vlamen” (Flemings) and “wallonen” (Walloons). In fact, this gives a hint for the representation of a central German perspective on Belgium during World War I in *Süddeutsche Monatshefte*. In political as well as in media debates there was a concentration on the Flemings as a seemingly Germanic people which had to be protected from French or rather Romanic influence and oppression in a multilingual state. Underlying concepts were based on ideas of divide and conquer, the fear of encirclement, and ultimately *völkische* and racist notions of nations and their structures.<sup>42</sup> Of course, this cannot be derived directly from the word cloud. Rather, it is based on prior knowledge originating from close readings of articles in other journals and from readings of research on Belgium during World War I. So it is unlikely but possible that the articles in *Süddeutsche Monatshefte* with their focus on Flemings and Walloons take a critical stance on this separation.

What helps, then, is the use of Voyant’s contexts tool<sup>43</sup> to examine the use of the single tokens. It is not absolutely necessary to read every single instance in this tool’s panel – in this case, for example, some hundred appearances of “wallon\*”, “flam\*”, “fläm\*”,<sup>44</sup> and so on would have been to checked. Yet even a cursory look at the results confirms the above suppo-

---

42 Cf. *Schaepdrijver, Sophie De*, Belgium, in: John Horne (ed.), *A Companion to World War I*, Chichester: Wiley-Blackwell, 2010, 391–393. For a more detailed presentation of German images of Belgium during World War I cf. *Bischoff, Sebastian*, *Kriegsziel Belgien: Annexionsdebatten und nationale Feindbilder in der deutschen Öffentlichkeit, 1914–1918*, Münster/New York: Waxmann, 2018.

43 Tools like this are usually known as Keyword in Context (KWIC). They generate a list of any instance of a word in question with a variable context of words on the left and right side of the given token.

44 Here, again, the lack of NER and tokenization is annoying for any instance of these words has to be searched separately.



sition: Indeed, a good part of the instances is concerned with a definite differentiation and separation of the Flemings and Walloons and with the attempt to highlight the Flemings as a Germanic people. At this point, the term “sprache” (language) in the word cloud comes into play. When using the contexts tool again, the results highlight the significant connection of language with the Flemings in the documents, whereby Flemish is presented as a suppressed language, which has to be protected and supported by the Germans.

Now this might lead to the presumption that instances of the aforementioned terms occur in the context of terms like “verwaltung” (administration), “regierung” (government, administration), and “politik” (politics, policy) and that there might be proposals in the articles for an occupational administration to the advantage of the Flemings. However, this is only correct in some way. In fact, the occurrences of “politik” refer to the foreign politics of the former Belgian government and the British government in most cases; and the occurrences of “regierung” refer to the domestic and foreign politics of the former Belgian government most often. Only the term “verwaltung” is used in the supposed sense. Does that mean that in the majority of cases the contributors of *Süddeutsche Monatshefte* argue in favor of administrative measures under military occupation of Belgium instead of an annexation? At least this is what Kauffmann supposes based on his close reading of the April 1915 issue on Belgium.<sup>45</sup>

But there is another problem with the term “verwaltung”. As in the example of the article on the Turkish straits mentioned above, a further examination of the results reveals that its appearance in the word cloud is bound to a good part to one single article on Belgium under German administration.<sup>46</sup> This does not necessarily mean that governmental and administrative questions in occupied Belgium are of minor importance than the word

---

45 Cf. K. Kauffmann, Wissensvermittlung, Kulturpolitik und Kriegspropaganda, 125.

46 Bissing, Friedrich Wilhelm Freiherr von, Belgien unter deutscher Verwaltung, in: *Süddeutsche Monatshefte* 12 (1915), 74–93. The same is true for the word “unterricht” (education, teaching) whose occurrence is bound to a good part to the article of Ziegeler, Jozef Haller van, Der mittlere Unterricht in Belgien, in: *Süddeutsche Monatshefte* 13 (1916), 605–616. This also points to a yet unmentioned problem with the underlying raw text files. Since there is no further annotation, page headers with the name of the author and/or the title of the article are always counted as a new token of a word and thus lead to erroneous results.

cloud – and Kauffmann in his article – suggest; then again, these topics probably are to be recognized on another level than words. One possibility could be the connection and arrangement of articles in the whole issue as discussed below.

## 2.2 Testing for hypotheses with dirty OCR

The discussion in the section above was primarily led by a deductive method, using Voyant as a means to reveal some of the foci of *Süddeutsche Monatshefte* and to go deeper until a point is reached where close reading seems to be the best option for further investigation. For this approach, no or less presuppositions were needed. However, as mentioned above, there are some strong assumptions about the corpus developed and published by Kai Kauffmann. Is it possible to reassess them with dirty OCR and Voyant? In the following section, some of these assumptions will be presented and possibilities to test them will be discussed.

(1) Kauffmann suspects at least a small increase of globalized horizons of comparison in German wartime journalism especially due to the entry of the USA and Japan into the war.<sup>47</sup> This assumption seemed to be easy to prove. Using Voyant's trends tool, I searched the document for "amerika\*" (America)<sup>48</sup> and "japan\*"<sup>49</sup> with the result depicted in figure 7. The columns indeed show a continuous occurrence of both terms throughout the corpus with some significant peaks. When asking for the reason of those peaks, the solution was somewhat obvious though disappointing for any height rested upon articles and main topics, which could have been identified easily with a close look on the table of contents. In this light, the advantage of digital analysis tools is only a faster result compared to counting the articles manually.

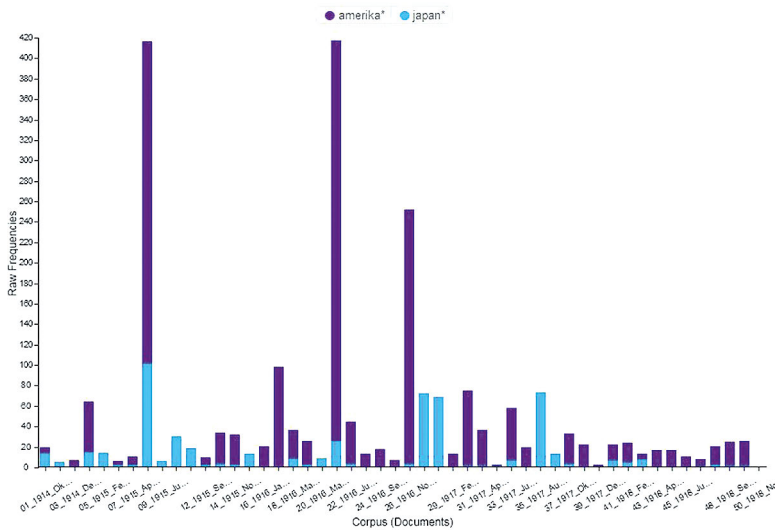
---

47 K. Kauffmann, *Wissensvermittlung, Kulturpolitik und Kriegspropaganda*, 120.

48 The search for "USA" and "Vereinigte Staaten" (United States) turned out to be of not even minor significance; for the reasons cf. the discussion on single terms and the problems with missing annotations above in section 2.1.

49 In the face of dirty OCR, I searched for instances of "japan\*" as well. But as there were only two results in the whole corpus, I left those instances out of account.

Fig. 7: Raw frequencies of “amerika\*” and “japan\*” in the whole corpus



Nevertheless, the created visualization goes beyond Kauffmann’s insight in some respect. On the one hand, it is a means for a more precise and striking presentation of what stays kind of vague in Kauffmann’s article.<sup>50</sup> On the other hand, it reveals a possible decline of globalized comparisons at the turn of 1916/1917 – at least there is no more actual focus on Japan or the United States. This is especially surprising in light of the United States’ entry into the war in April 1917, an event producing lots of media coverage, one would think. Is it the monthly publication frequency of the journal that makes it difficult to react to even such an important event? Or do events and circumstances at the home front become more important?<sup>51</sup> Further investigation in this respect

50 Moreover, it calls attention to the nearly always quite careless handling of quantifiable statements in the philologies. Of course, quantification in a statistical sense is no adequate option in any case – be it because an exact value is of no further explanatory power or be it because no digitized corpus is available. But in some instances, digital tools definitely help to underline insights which would otherwise be nothing more than unverifiable claims. On the contrary, statistical outcomes are far away from being self-explanatory. For a further discussion of these epistemological questions of Digital Humanities cf. B. Rieder/T. Röhle, *Digital Methods*, 71–79.

51 There are some hints for this supposition because many of the later wartime issues of *Süddeutsche Monatshefte* deal, for example, with German agriculture (July 1917), German

is needed, but the potential of digital tools to move from an initial question to a quite different one, which was out of sight before, becomes obvious.

(2) Another important finding of Kauffmann is the outstanding position of historians within the contributors of *Süddeutsche Monatshefte*. In contrast, philologists, or especially natural scientists, play only a marginal role, if at all.<sup>52</sup> Follow-up questions could aim at differences in the use of distinctive terms or patterns of arguments depending on the academic profession of the contributors. Once again, it is not dirty OCR that causes problems but the lack of any further annotation, in this case regarding metadata on the contributors and their academic profession – a gap that should be closed for further work with the digital corpus and its analysis with digital tools, let alone the publication of a digital edition of the documents –, which makes the examination difficult.<sup>53</sup> Of course, also in this case the selection and formation of a subcorpus could be done; but, in the end, this is a whole new stage of work consuming lots of time and therefore costs – work I have not done yet so I am unable to present any results.

(3) Finally, there is an essentially theoretical perspective on the *form* of periodicals in Kauffmann's approach, asking for their special ability to build up comparisons – or opportunities to compare – due to the arrangement of their material.<sup>54</sup> Almost any journal brings together articles by different authors with the same or different views on the same or related topics,<sup>55</sup> thus enabling the reader not only to accumulate the knowledge but also to compare between those views. So journals can stabilize or destabilize existing opinions or even formations of discourse. Here, the focus is not so much on

---

social democracy (November 1917), German industry (March 1918), or German workers (January 1918).

52 K. Kauffmann, *Wissensvermittlung, Kulturpolitik und Kriegspropaganda*, 127.

53 The same problem arises, by the way, when asking for differences between different factional and fictional genres. Even when testing for the ability of a tool to differentiate between, genres a comparative corpus generated by a human being is needed; cf., for example: Allison, Sarah et al., *Quantitative Formalism: an Experiment*, accessed: 14.05.2019, <https://litlab.stanford.edu/LiteraryLabPamphlet1.pdf>.

54 K. Kauffmann, *Wissensvermittlung, Kulturpolitik und Kriegspropaganda*, 124–26.

55 Exceptions are, for example, journals that are not only edited but written to a large extent by only one person like Karl Kraus' "Die Fackel" or Maximilian Harden's "Die Zukunft". Other exceptions might be periodicals with a powerful editorial board and a rigorous political agenda.

the content of single articles but on the arrangement of an ‘ensemble of texts’<sup>56</sup> and therefore on the media preconditions of practices of comparison, their effects on practices of comparison and the modalities of the production of knowledge. In fact, this is by the far most difficult aspect to test with Voyant. To explain the underlying concept in more detail and to present the problems with Digital Humanities in this case, I will focus on the July 1915 issue of *Süddeutsche Monatshefte* which has an emphasis on “Rußland von Innen” (“Russia from the inside”). First, I will show some of the results of Voyant when processing this issue to then compare them with central insights gained by close reading.

Again, I started with a word cloud whereby I deleted any instances of “deutsch-” (German) and “russ-” in addition to the OCR errors to reveal more of the things beyond the expectable (see figure 8). Though there are a lot of remarkable objects in this cloud to focus on,<sup>57</sup> I will concentrate on one aspect only: women and the question of gender.

In the historical context of the document, it seems evident that there is a close relationship between women – depicted on the upper left side of the cloud (“frauen”) – and family-related words like “kinder” (children) and “sohn” (son). A good way to test supposed correlations with Voyant is the employment of its collocates tool which shows words found in proximity of a keyword. As figure 9<sup>58</sup> reveals, there is indeed a correlation between women

---

56 K. Kauffmann, *Wissensvermittlung, Kulturpolitik und Kriegspropaganda*, 124.

57 Especially the central position of the word “jüdischen” (Jewish) would need closer attention. While in this case it is obviously caused by a longer article on the situation of the Jewish proletariat in Russia, it is remarkable that Jews are the only ethnic or religious group that appears in figure 5, too. Is this focus on the Jews a result of antisemitism? Or is it because of the encounter with (orthodox) Jews in the occupied territories in the East? Do the contributors of *Süddeutsche Monatshefte* discuss possibilities to help the mainly poor and suppressed Jews in tsarist Russia or do they focus on measures to separate the Jews? For a general discussion of these questions cf., for example, *Zechlin, Egmont, Die deutsche Politik und die Juden im Ersten Weltkrieg*, Göttingen: Vandenhoeck u. Ruprecht, 1969; *Angress, Werner T., Das deutsche Militär und die Juden im Ersten Weltkrieg*, in: *Militärgeschichtliche Mitteilungen* 19 (1976), 77–146; and *Hoffmann, Christhard, Between Integration and Rejection: The Jewish Community in Germany, 1914–1918*, in: John Horne (ed.), *State, Society and Mobilization in Europe during the First World War*, Cambridge: Cambridge University Press, 1997, 89–104.

58 The context is restricted to five words on each side; depicted are only those collocates which appear at least three times.

Fig. 8: Word cloud of the July 1915 issue of Süddeutsche Monatshefte with some additions to the stop list



and family, as “mutter” (mother) and “kinder” are among the top collocations. Not very surprising, too, is the occurrence of “männer” (men) in the list because talking about gender issues was – and still is – almost always talking about differences between the sexes and about their relationship. So what really draws attention are the words “sachalin” (Sakhalin), “revolution” and “beteiligung” (participation).

Fig. 9: Collocations of “frau\*” in the July 1915 issue of Süddeutsche Monatshefte

<input type="checkbox"/>	Term	Collocate	Count (context)
<input type="checkbox"/>	frau*	revolution	5
<input type="checkbox"/>	frau*	männer	5
<input type="checkbox"/>	frau*	mutter	4
<input type="checkbox"/>	frau*	kinder	4
<input type="checkbox"/>	frau*	beteiligung	4
<input type="checkbox"/>	frau*	sachalin	3

The occurrence of “sachalin” is obviously due to the imprint of a report by Anton Čechov on the penal colony on the island Sakhalin.<sup>59</sup> But why does it feature women? The article includes a section on “Die Frauen und Kinder von Sachalin” (“The Women and Children of Sakhalin”) where Čechov talks about women who committed a “crime, almost exclusively murder, which rests upon love affairs and family conflicts”<sup>60</sup> and women who followed their sentenced men to Sakhalin. Without any income opportunity, they would sooner or later engage in prostitution.<sup>61</sup> Yet this is the result of a close reading of the article while the digital tools did not reveal the exact same insights. Even if tools that help to reach comparable results exist, traditional forms of reading are the method of choice when it comes to analyzing a relatively short article like the one in question.

But what is revealed by this is the possible connection to two other articles of this issue: to an article by Adolf Dirr on “Die Russin”<sup>62</sup> (“The Female Russian”) and to an article by Nadja Straßer on “Die russische Frau in der Revolution” (“The Russian Woman in the Revolution”).<sup>63</sup> The titles of both texts already reveal their interest in gender issues but the constellation is remarkable in more than one respect. On the surface, a female and a male author simply voice their opinion on related topics in the two articles. While Dirr explicitly points to the fact that he is writing from a male standpoint,<sup>64</sup> there is an editor’s note above Straßer’s article which declares his contribution to represent the European view and hers to be the view of a “liberal Russian woman”<sup>65</sup> – by the way, a Jewish feminist who moved to Vienna in the late 1890s and lived in Berlin at the time of publication of

---

59 *Tschechow, Anton*, Die Gefängnisinsel Sachalin, in: *Süddeutsche Monatshefte* 12 (1915), 701–710.

60 *Ibid.*, 708: “[...] Verbrechen, fast ausschließlich Mord, [die] auf Liebesaffären und Familienzwistigkeiten beruhen [...]”.

61 *Ibid.*, 709.

62 *Dirr, Adolf*, Die Russin, in: *Süddeutsche Monatshefte* 12 (1915), 588–596.

63 *Straßer, Nadja*, Die russische Frau in der Revolution, in: *Süddeutsche Monatshefte* 12 (1915), 647–652.

64 Cf. A. *Dirr*, Die Russin, 588. For biographical information on Adolf Dirr cf. *Öhrig, Bruno*, Adolf Dirr (1867–1930): Ein Kaukasusforscher am Münchner Völkerkundemuseum, in: *Münchner Beiträge zur Völkerkunde* 6 (2000), 199–234.

65 *N. Straßer*, Die russische Frau in der Revolution, 647: “Wir nehmen an, daß es für unsere Leser Wert hat, nachdem sie in Dr. Dirrs Aufsatz den europäischen Maßstab angelegt

the article.<sup>66</sup> Thus cultural and gender aspects of the authors are interwoven, and the article is presented as an offer for comparison by the editor of *Süddeutsche Monatshefte* in light of different aspects of authorship.<sup>67</sup>

What do both address in detail? Dirr, while praising the Russian woman for “being more natural, less spoiled by culture” than the West European woman – thus connecting stereotypes about women and Russians as creatures of nature – focuses on her character, which he finds nevertheless to be “vain, empty, saucy, haughty, cheeky”.<sup>68</sup> Straßer, in contrast, emphasizes the “spontaneity and certainty” (“Ungezwungenheit und Sicherheit”) of the Russian woman that makes her tend to a comradely relationship to men and to revolutionary movements.<sup>69</sup> Most instances of the word “beteiligung” (participation) in the results of the collocates tool can be found in this text: It is the participation of women in the revolutionary action in 1905 in Russia.

Now, what might that mean for readers of *Süddeutsche Monatshefte* – predominantly male members of the educated bourgeoisie? Though Straßer’s article is full of sympathy for female and social insurrection, it is framed by Dirr’s article and Čechov’s report. In this regard, when it comes to comparing the three texts, Straßer’s perspective might be nothing more for contemporary readers than an affirmation of Dirr’s chauvinistic view of Russian women, who, finally, end up in Čechov’s penal colony for their tendency to insubordination. This is clearly not the result of the intention of the contributors or the message of the single articles. Instead, it is due to the compilation and arrangement of these articles in the same issue of *Süddeutsche Monatshefte*.

---

sahen, nun auch den spezifisch russischen Standpunkt vertreten zu sehen, indem wir einer freiheitlichen Russin das Wort geben.”

66 Cf. Schmidt, Birgit, ‘Die Frauenpflichtlerin’ – Zur Erinnerung an Nadja Strasser, in: *Aschenas* 16 (2006), 229–259.

67 The editor at this time was Paul Nikolaus Cossmann; cf. Selig, Wolfram, Paul Nikolaus Cossmann und die Süddeutschen Monatshefte von 1914–1918: Ein Beitrag zur Geschichte der nationalen Publizistik im Ersten Weltkrieg, Osnabrück: A. Fromm, 1967.

68 Dirr, “Die Russin,” 592: “Eitel, leer, naseweis, hochmütig, vorlaut [...]”

69 Straßer, “Die russische Frau in der Revolution,” 649. Nevertheless, she reproduces stereotypical representations of Russians as well when calling them humans of emotion instead of action (“nicht Tat-, sondern Gefühlsmensch”) and “half wild and primitive” (“halbwild und primitiv”); *ibid.*, 647.



It is this potential of journals to produce meanings beyond single articles by addition of knowledge and claims and by comparison of knowledge and claims what makes them special and interesting for research. But these procedures of writing, editing, and reading rest upon structures that are hard to be detected with tools like those made available by Voyant – and, perhaps, rest upon structures so closely connected with human understanding that it is inevitable to return to close reading, even if one has started with distant reading.<sup>70</sup>

### 3. Conclusion

Without any doubt, dirty OCR is not appropriate for the production of robust and final results of any research. Too many errors make it impossible to detect any occurrence of certain words. Moreover, the lack of any further annotations, lemmatization or named entity recognition disturbs the quantifiable statistical outcome. Nevertheless, it works well enough when it comes to prove hypotheses about larger trends in a given corpus or to reveal something more about larger patterns. Therein lies the greatest potential of dirty OCR. Of course, when working with digital tools some adjustments have to be made – an addition of the stop list, for example, – and greater attentiveness to the (formal) structure of the digitized corpus is needed in contrast to a ‘clean’ corpus. But these problems are manageable to a great extent.

This is true, too, for the work with a web-based tool like Voyant. Publications like those of Franco Moretti and others have shown much more of the potentials of Digital Humanities, but their authors are conducting their research in a nearly perfect scientific environment with computer scientists who are available any time to refine the abilities of tools again and yet again. For those of us who lack this close connection with computer experts, free accessible tools – alongside clean or dirty OCR – are at least an option for starting research.

---

70 Cf. for example Stefan Scherer and Claudia Stockinger, “Archive in Serie: Kulturzeitschriften des 19. Jahrhunderts,” in *Archiv/Fiktionen: Verfahren des Archivierens in Literatur und Kultur des langen 19. Jahrhunderts*, eds. Daniela Gretz and Nicolas Pethes (Freiburg, Berlin, and Wien: Rombach, 2016), 268, with their claim that digital tools can not reproduce the seriality of periodicals so that close reading is irreplaceable in periodical studies.

Whether dirty OCR and web-based tools can be of help in everyday research or not, largely depends on the research questions, of course. Voyant proved to be a valuable tool, especially when working with it in a mere deductive way with no or less presuppositions about the corpus. On the one hand, Voyant's word clouds made it possible to go deeper and deeper into the corpus while creating new questions or refining preliminary assumptions. On the other hand, as the discussion above has shown, the effectiveness of Voyant and probably of any digital tool drops when the research focuses on a smaller number of articles.

In this regard, the production of documents with dirty OCR is probably most suitable for mid-scale corpora with a good rate of production time and possible research results – for the larger a corpus is, the more time is needed for its digitization. This time might be wasted when the digitization is not finished in accordance with (digital) editorial standards; the smaller a corpus is, on the contrary, the less will be the advantage of using digital tools.

In an academic world that is driven more and more to be project-based, which means the probable outcome of research should at best be known before any work has started, provisional digitized documents could definitely serve as a basis for the verification and/or development of hypotheses. Though not necessarily adoptable for publication, the results might at least be a basis for better-grounded assumptions in project proposals.

## Bibliography

- Allison, Sarah et al., Quantitive Formalism: an Experiment, accessed: 14.05.2019, <https://litlab.stanford.edu/LiteraryLabPamphlet1.pdf>.
- Angröss, Werner T., Das deutsche Militär und die Juden im Ersten Weltkrieg, in: Militärgeschichtliche Mitteilungen 19 (1976), 77–146.
- Anthony, Laurence, A Critical look at software tools in corpus linguistics, in: Linguistic Research 30 (2013), 141–61.
- Archer, Dawn, Data Mining and Word Frequency Analysis, in: Gabriele Grifflin/Matt Hayler (eds.), Research Methods for Reading Digital Data in the Digital Humanities, Edinburgh: Edinburgh University Press, 2016, 72–92.
- Bischoff, Sebastian, Kriegsziel Belgien: Annexionsdebatten und nationale Feindbilder in der deutschen Öffentlichkeit, 1914–1918, Münster/New York: Waxmann, 2018.

- Bissing, Friedrich Wilhelm Freiherr von*, Belgien unter deutscher Verwaltung, in: *Süddeutsche Monatshefte* 12 (1915), 74–93.
- Collier, Patrick*, What is Modern Periodical Studies?, in: *Journal of Modern Periodical Studies* 6 (2015), 107–108.
- Cordell, Ryan*, ‘Q i-jtb the Raven’: Taking Dirty OCR Seriously, in: *Book History* 20(2017), 194–95.
- Dirr, Adolf*, Die Russin, in: *Süddeutsche Monatshefte* 12 (1915), 588–596.
- Hahn, Carolin*, Forschung benötigt Infrastrukturen: Gegenwärtige Herausforderungen literaturwissenschaftlicher Netzwerkanalysen, in: Toni Bernhart et al. (ed.), *Quantitative Ansätze in den Literatur- und Geisteswissenschaften: Systematische und historische Perspektiven*, Berlin and Boston: de Gruyter, 2018, 315–34.
- Hoffmann, Christhard*, Between Integration and Rejection: The Jewish Community in Germany, 1914–1918, in: John Horne (ed.), *State, Society and Mobilization in Europe during the First World War*, Cambridge: Cambridge University Press, 1997, 89–104.
- Holley, Rose*, How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs, in: *D-Lib Magazine* 15 (2009), <https://doi.org/10.1045/march2009-holley>.
- Horne, John/Kramer, Alan*: *German Atrocities, 1914: A History of Denial*, New Haven: Yale University Press, 2002.
- Jannidis, Fotis/Kohle, Hubertus/Rehbein, Malte (eds.)*, *Digital Humanities: Eine Einführung*, Stuttgart: Springer, 2017
- Kauffmann, Kai*, Wissensvermittlung, Kulturpolitik und Kriegspropaganda: Thesen zur Kriegspublizistik der deutschen Rundschauzeitschriften 1914–1918, in: Olivier Agard/Barbara Beßlich (eds.), *Krieg für die Kultur? Une guerre pour la civilisation? Intellektuelle Legitimationsversuche des Ersten Weltkriegs in Deutschland und Frankreich (1914–1918)*, Berlin: Peter Lang, 2018, 113–128.
- Latham, Sean, Scholes, Robert*, The Rise of Periodical Studies, in: *PMLA* 121 (2006), 517–531.
- Nicholson, Bob*, Counting Culture; or, How to Read Victorian Newspapers from a Distance, in: *Journal of Victorian Culture* 17 (2012), 238–246.
- Öhrig, Bruno*, Adolf Dirr (1867–1930): Ein Kaukasusforscher am Münchner Völkerkundemuseum, in: *Münchner Beiträge zur Völkerkunde* 6 (2000), 199–234.

- Reventlow, Ernst*, Die Frage der türkischen Meerengen und ihre Entwicklung, in: *Süddeutsche Monatshefte* 14 (1917): 432–66.
- Riddell, Allen Beye*, How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models, in: Lynne Tatlock/Matt Erlin (eds.), *Distant Readings: Topologies of German Culture in the long nineteenth Century*, New York: Rochester, 2014, 91–113.
- Rieder, Bernhard/Röhle, Theo*, Digital Methods: Five Challenges, in: David M. Berry (ed.), *Understanding Digital Humanities*, New York: Palgrave Macmillan, 2012, 67–84.
- Roff, Sandra*, From the Field: A Case Study in Using Historical Periodical Databases to Revise Previous Research, in: *American Periodicals* 18 (2008), 96–100.
- Schaepdrijver, Sophie De*, Belgium, in: John Horne (ed.), *A Companion to World War I.*, Chichester: Wiley-Blackwell, 2010, 386–402.
- Schmidt, Birgit*, 'Die Frauenpflichtlerin' – Zur Erinnerung an Nadja Strasser, in: *Aschkenas* 16 (2006), 229–259.
- Selig, Wolfram*, Paul Nikolaus Cossmann und die Süddeutschen Monatshefte von 1914–1918: Ein Beitrag zur Geschichte der nationalen Publizistik im Ersten Weltkrieg, Osnabrück: A. Fromm, 1967.
- Strange, Carolyn et al.*, Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers, in: *Digital Humanities Quarterly* 8 (2014), <http://www.digitalhumanities.org/dhq/vol/8/1/000168/000168.html>.
- Straßer, Nadja*, Die russische Frau in der Revolution, in: *Süddeutsche Monatshefte* 12 (1915), 647–652.
- Tschechow, Anton*, Die Gefängnisinsel Sachalin, in: *Süddeutsche Monatshefte* 12 (1915), 701–710.
- Zechlin, Egmont*, Die deutsche Politik und die Juden im Ersten Weltkrieg, Göttingen: Vandenhoeck u. Ruprecht, 1969.
- Ziegeler, Jozef/Haller van*, Der mittlere Unterricht in Belgien, in: *Süddeutsche Monatshefte* 13 (1916), 605–616.



## Challenging the *Copia*

# Ways to a Successful Big Data Analysis of Eighteenth-Century Magazines and Treatises on Art Connoisseurship<sup>1</sup>

---

Joris Corin Heyder

### Introduction

Being able to compute big data in a relatively short span of time is one of the greatest advantages of DH projects. While it is almost impossible to critically read and analyze a vast corpus of c. 550 titles with an average of 300 pages, a fully digitized corpus ideally combined with linguistic indexing is relatively easy to examine in respect of a particular semantic field and structure, specific notions, word co-occurrences, and more. However, what if the corpus is not available as machine-encoded text, but only in form of text-images from scanned documents in varying, sometimes very poor qualities? As conventional optical character recognition (OCR) has a high error rate particularly in cases of the older literature used in my project, where it is for instance necessary to differentiate for example between the “long s” (ſ) and “f” characters, the machine encoded results are often hardly useable for further operations. Particular difficulties arise out of the diversity of the underlying material that comprises connoisseurial and philosophical French and German treatises published between 1670 and 1850 as well as magazines like the *Mercure de France* or the *Teutsche Merkur* of the same period. This is true on different levels, first, the quality and resolution of the scanned text-images, second,

---

<sup>1</sup> This article has been written within the framework of the Collaborative Research Center SFB 1288 “Practices of Comparing. Changing and Ordering the World”, Bielefeld University, Germany, funded by the German Research Foundation (DFG), subproject C 01, “Comparative Viewing. Forms, Functions and Limits of Comparing Images”. My heartfelt thanks to Julia Becker for proofreading my manuscript for any linguistic and stylistic errors.

the typographical wide range from German black letter print to French Antiqua, and, third, the high quantity of texts in magazines that has been written on other topics than art connoisseurship. In this paper I am seeking to propose a heuristic how to tackle both, the diverse conditions of the once digitized material as well as the potentials offered by open source corpus analysis toolkits, such as AntConc.<sup>2</sup> I will argue for a combination of a “quick and dirty” approach to mass digitization with a complementary reflecting and recontextualizing close reading.

## 1. Research design and problems in collecting a data corpus

When I started my project on the practices of comparing in eighteenth-century French and German art connoisseurship by first gathering a huge amount of resources from well-known and less well-known connoisseurial treatises, exhibition guides, magazines, etc., from time to time, I had to think about the etymological roots of the word “copy” that originates from the Latin *copia*, which meant “abundance” or “richness of material”. In medieval manuscript culture the French term “copie” has been extended to its current meaning as “duplication, imitation, or reproduction.” However, both sides of the coin are present in looking at the plentiful material which I had to tackle after only a short span of time. On the one hand, I had the great opportunity to find even the rarest printed material in searching engines like Gallica<sup>3</sup> or the digital library of the Bayerische Staatsbibliothek.<sup>4</sup> On the other hand, this *copia* (in the sense of “abundance”) was just too vast to scour for the examples I was seeking to find, namely, explicit and implicit reflections on the role of comparing visually in art connoisseurship. The many thousands of scanned pages, or to be more precise, digital copies of distinctive genera of publications were a curse and a savior at the same time. In pre-digital times, before the tempting offer of a veritable Babylonian

---

2 The software can be downloaded here: <https://www.laurenceanthony.net/software.html> [accessed: 08.05.2019].

3 <https://gallica.bnf.fr/accueil/fr/content/accueil-fr?mode=desktop> [accessed: 08.05.2019].

4 [https://www.digitale-sammlungen.de/index.html?c=sammlungen\\_kategorien&l=de](https://www.digitale-sammlungen.de/index.html?c=sammlungen_kategorien&l=de) [accessed: 08.05.2019].

library,<sup>5</sup> ‘natural’ limitations were stipulated by the possessions of the libraries themselves. Of course, Gallica represents nothing else than the immense possessions of the *Bibliothèque nationale de France* and 270 other French institutions like *Bibliothèques municipales* or *Instituts de recherche*,<sup>6</sup> but everyone who ever worked in such institutions knows how difficult it sometimes can be to receive each volume one is interested in. Processes like ordering the copy, waiting for the keeper to bring it, as well as institutional frames like delimited opening hours, restricted use, and so on, limit the opportunities to see a great amount of material. Therefore, research designs are usually reduced to a certain corpus of texts and this corpus may grow and change over time, but this only happens within an adequate framework. The contrary is believed to be true with digital material that promises the user a prolific disposability of every thinkable relevant source from all over the world. This, of course, is no more than a phantasmagory. However, the potential to make every thinkable information instantaneously available led Paul Virilio (1932–2018) to the idea of the “digital ages as ‘the implosion of real time’”.<sup>7</sup> Silke Schwandt has taken the idea a little further and asked: “Do we live in a time of ‘eternal now’?”<sup>8</sup> What almost brings a transcendent taste to the discussion is yet worth being applied to a concrete example. A time of ‘eternal now’ would mean, amongst other things, that – from a digital perspective – a treatise written by Roger de Piles (1635–1709) has to be considered as present as the latest breaking news. Always entwined into a multiplicity of time layers – for instance their time of origin, their proper time or their historical time –, the tremendously synchronicity of information in itself epitomizes a huge problem. The user might get overwhelmed by the potential presence of the ‘eternal now’ but has to choose after all, which “now” comes first. Even worse, some of the ‘nows’ are more present than others, some even remain invisible: One could suggest that the claimed universality of resources will never become more than a chimera.

---

5 Cf. *Borges, Jorge Luis*, *The Library of Babel*, in: *Ficciones*, transl. by Anthony Kerrigan et al., New York: Grove Press, 1962.

6 Although ‘Gallica’ is labeled as a service of the *Bibliothèque nationale de France*, its partners are distributed over all regions of France, cf. <https://gallica.bnf.fr/html/decouvrir-nos-partenaires> [accessed: 23.04.2019].

7 *Schwandt, Silke*, *Looking for ‘Time’ and ‘Change’: Visualizing History in the Digital Age* (draft version, forthcoming), 12.

8 *Ibid.*



Thus, working with digital texts needs clear frameworks, too. Often, they are set more or less hazardingly by the resources themselves: How easy are they to find? Is the access to the digitized resources restricted? Is it possible to download the files, and if yes, is the quality of the scans sufficient for the needs of OCR or not? Of course, these limitations should not and cannot be decisive to handle the flood of information. How, then, is it possible to cope with the *copia*? An answer to this allegedly simple question can potentially be found in Petrarch's request to pledge yourself to *sufficiencia* ('moderation'). In chapter 43 entitled "De librorum copia" of his work *De remediis utriusque fortunae*<sup>9</sup> the allegorical figures *Gaudium* ('Joy') and *Ratio* ('Reason') dispute the challenges of the *copia* of books a long time before the invention of the letter press.<sup>10</sup> It was not the excess of accessible information but rather the unsuspecting use that motivated Petrarch to formulate his media critical remarks. He warns against the growing quantity of information and recommends being moderate in its use.

## 2. Digitizing the corpus

"Tene mensuram"<sup>11</sup> ('be moderate') perhaps should also have been the motto in the beginning of my project, but it was not. Instead, I optimistically started the research – as has already been mentioned – by bringing together hundreds of PDF files with more or less relevant content in expectation of a quick OCR process. My naïve optimism was guided by the experiences with commercial OCR software that I used for almost all of my scanned secondary literature. As the average of the results appeared to be valuable, I presumed that an improved OCR engine developed by our IT-team would produce comparable results with regard to the texts I was working on. The idea was to establish a quick and dirty approach<sup>12</sup> that allowed me to find relevant

---

9 A fully digitized version of c. 1490 probably printed by the editor Heinrich Knoblochtzger can be found here: <http://mdz-nbn-resolving.de/urn:nbn:de:bvb:12-bsb11303461-1> [accessed: 21.04.2019].

10 Siegel, Steffen, *Tabula: Figuren der Ordnung um 1600*, Berlin: Akademie-Verlag, 2009, 31.

11 Motto of German Emperor Maximilian I (1493–1519).

12 Kohle, Hubertus, *Digitale Bildwissenschaft*, Glückstadt: Verlag Werner Hülsbusch. Fachverlag für Medientechnik und -wirtschaft, 2013, 37–38. For a full digitized version, cf. <http://archiv.ub.uni-heidelberg.de/artdok/volltexte/2013/2185> [accessed: 08.04.2019].

words, co-occurrences, as well as characteristic semantic markers for comparisons, such as “plus ancienne/plus jeune”, “copié d’après”, “pareillement/different”, “moins avantageuse que”, etc. I intuitively estimated that it would be sufficient to reach an average text recognition of c. 80 percent to be able to run first tests with applications like *AntConc*. Only much later I learned that in machine learning based projects that use handwritten text recognition like Transkribus<sup>13</sup> a character error rate (CER) of less than five percent on average could be achieved.

Several months went by until the IT-team could first establish an optimized version of the free OCR software *Tesseract*.<sup>14</sup> It implemented specific requirements, for instance, the recognition of old font types like Gothic/Fraktur fonts as well as layout detection.<sup>15</sup> However, a major problem consisted in the scans itself: As Sven Schlarb of the Österreichische Nationalbibliothek (Austrian National Library) has shown in a talk at the final IMPACT conference (Improving Access to Text)<sup>16</sup> in 2010, an optimal scan is crucial to obtain optimum results. Of course, besides the font question, typical challenges of historical material appeared in most of the PDFs like warped book pages, curved text lines, different print intensities, distortions and contaminants, handwritten annotations, complex layouts, and of course time-specific orthography.<sup>17</sup> From a digital analytical perspective the PDFs had to be parsed. One item or one ‘now’ – the PDF – had not only to be dissolved into images or n-‘nows’ but also by help of a functional extension parser into a multiplicity of elements, like the text block, page numbers, characteristic layout features, and so forth. These items, then, had to be enhanced by different steps like border detection, geometric correction (‘unwarping’), as well as binarization, i. e., the conversion of a picture into black and white. Although those processes are automatable to a certain degree, they maybe take the longest time. It is worth illustrating that with the different outcomes gained within the digitization process: A characteristic quality acces-

13 Cf. <https://transkribus.eu/Transkribus/> [accessed: 13.06.2019].

14 [https://en.wikipedia.org/wiki/Tesseract\\_\(software\)](https://en.wikipedia.org/wiki/Tesseract_(software)) [accessed: 08.04.2019].

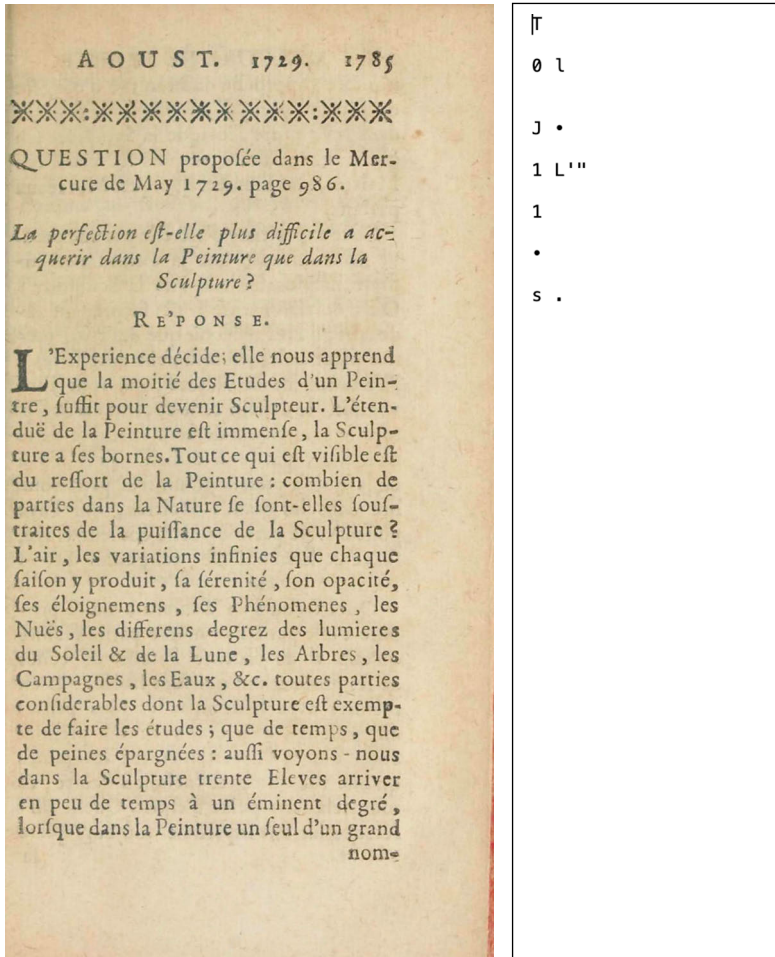
15 For a precise description of these operations, cf. the paper by Patrick Jentsch and Stephan Porada in this volume.

16 <https://impactocr.wordpress.com/2010/05/07/an-overview-of-technical-solutions-in-impact/> [accessed: 08.04.2019].

17 <https://impactocr.wordpress.com/2010/05/07/an-overview-of-technical-solutions-in-impact/> [accessed: 08.04.2019].

sible at Gallica, present in an example taken from the magazine *Mercur de France* from August 1729 (ill. 1a) promises prosperous results.

Ill. 1a & 1b

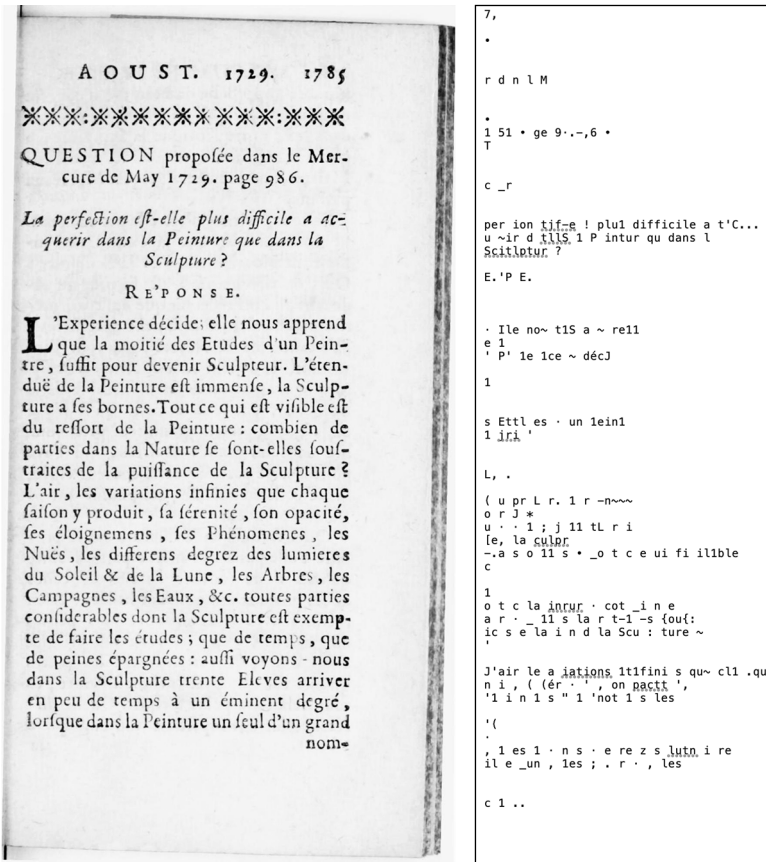


The scan is clearly legible, the page is not warped, the page layout appears to be not too difficult to “read”. A first test with a commercial OCR performing PDF-reader shows, however, the limited capacity of such applications ever more drastically to us. In fact, by using the OCR tool with the options

“French” and “dpi dissolution 600” (ill. 1b) and by transferring the PDF to .txt-format afterwards zero result is achieved.

The effects of a first binarization (ill. 2a) demonstrate two things: first, an optimized image enormously increases the success of the ‘reading’ process, and, second, despite such achievements the effects appear to remain difficult to optimize (ill. 2b).

Ill. 2a & 2b



With the OCR function of the commercial PDF reader at least (and of all things) the word “difficile” was recognizable for the machine. Keeping those unsuccessful data in mind, the results of the optimized Tesseract OCR

(ill. 3a–c) go far beyond any expectations. But this is only true for anyone knowing the enormous difficulties visible in illustrations 1b and 2b. When I studied the outcomes of our first OCR tests I was shocked. Even if many words (ill. 3a), as for instance “Aoust” or “Peinture” in line one and six, were readable, others like “visible” in line thirteen became illegible because of the disassembling of the readable and the unreadable syllables and/or signs into two or more parts: “vi ble”.

### Ill. 3a & 3b & 3c

Mercure_de_France_-_1729_Août... X	Mercure_de_France_-_1729_Août... X	Mercure_de_France_-_1729_Août... X
1 A O U S T. 1719. 1735	1 A O U S T. 1929. 1785	1 A O U S T. 1729. 1786
2 :xxxxxxxxxxxxxxxxxxxx	2 QUES TION proposée dans le Mer-	2 QUES TION proposée dans le Mer-
3 QUE S T I Ø N proposée dans le Mer-	3 ; cure de May 1729. page 936.	3 QUESTION proposée dans la Mer-
4 cure de May 1729. page 98 6.	4 P	4 cure de May 1729. page 986.
5 L4 Perfg&i0 F -8/! plu: dij ctic  :zc i	5 ;	5 La perfection eff-elle plus difficile à ac-
6 querir dan la Peinture que dans la	6 le perfection est-elle plus difficile a ac-	6 querir dans la Peinture que dans la
7 Sculpture ?	7 l guerir dans la in que dans la	7 Sculpture ?
8 RE PONSE.	8 Sculpture:	8 Reronn&e.
9 Experience décide; elle nous apprend	9 ?	9 Experience décide; elle nous apprend
10 l que la. moitié des Etudes d un Pein-	10 k REeronn&e.	10 IL que la moitié des Etudes d'un Pein-
11 tre , fuer pour devenir Sculpteur. L'éten-	11 l Experiencec&e; elle nous apprend	11 tre, fuffit pour devenir Sculpteur. L'éten-
12 dué de la Peinture e imenne , la Sculp-	12 è que la moitié des Etudes d'un Pein-	12 dué de la Peinture est imenne, la Sculp-
13 ture & fes bornestout ce qui c&v i ble e :	13 è de la Peinture est imente , la Sculp-	13 ture a fes bornes.Tout ce qui est vifible est
14 du reffort de la Peinture : combien de	14 ture a fes bornes.Tout ce qui est vifible est	14 du reffort de la Peinture : combien de
15 parties dans la Nature fe font elles couf-	15 du reffort de la Peinture : combien de	15 parties dans la Nature fe font-elles fauf=
16 trates de la pui ance de la Sculptur&c	16 - parties dans la Nature fe font-elles fauf=	16 l traies de la puiffance de la Sculpture e
17 L'air, les variations in nies que chaque	17 traites de la puiffance de la Sculpture e	17 L'air, les variations infinies que chaque
18 _ faifon y produit , (a érentité , (on opacité,	18 l L'air, les variations infinies que chaque	18 _ faifon y produit , (a érentité , fon opacité,
19 fes éloignemens , fes Phénomènes ), les	19 : faifony produit, (a érentité , fon opacité,	19 [fes éloignemens , fes Phénomènes , les
20 Nués , les dihc ns degrez des lumieres	20 [fes éloignemens , fes Phénomènes les	20 Nués , les differens degrez des lumieres
21 du Soleil & de la Lune , les Arbres , les	21 l	21 du Soleil & de la Lune , les Arbres , les
22 Campagnes , les Eaux , &c. toutes parties	22 l tre , fuffit pour devenir Sculpteur. L'éten-	22 Campagnes , les Eaux , &c. toutes parties
23 con dcrables dont la Sculpture exemp-	23 Nués , les differens degrez des luinres	23 confiderables dont la Sculpture est exemp-
24 te de faire les études ; que de temps , que	24 du Soleil & de la Lune , les Arbres , les	24 te de faire les études ; que de tems , que
25 de peines épargnées : auffi voyons - nous	25 Campagnes , les Eaux , &c. toutes parties	25 de peines épargnées : auf-! voyons - nous
26 dans la Sculpture trente Eleves arriver	26 goniderables dont la Sculpture est exemp-	26 dans la Sculpture trente Eleves arriver
27 \	27 te de faire les études ; que de temps , que	27 en peu de temps à un éminent degré,
28 _ - I - :	28 de peines épargnées: auffi voyons - nous	28 lorsque dansla Peinture un feul d'un grand
29 en peu de temps a un emment degré,	29 dans la Sculpturetrente Eleves arriver	29 nome
30 lorfque dans la Peinture un feul d un grand	30 en peu de temps à un éminent degré,	
31 nom e	31 - lorfquedans la Peinture un feul d'un grand	
	32 l nome	
	33 Ta	
	34 j0-Vu	
	35 AS	
	36 œ	
	37 EE	

What was even worse was the recognition of cursive letters; the disintegration of the word “perfection” – that forms part of line five – into the heap of signs: “Perfg&i0” ensued a concrete problem for the spotting of words. This is true because one would only gain a decisive outcome by considering the word stem “Perf” that is far too vague for a target-oriented research. Better results could be achieved by a second, revised Tesseract-version (ill. 3b).

While the italics caused fewer problems, as the correctly recognized example “perfection” proves, also the word “vifible” is now represented as one integral word. Here, the long “s” is ‘read’ as “f”, which is a common but manageable problem in the underlying text genera of eighteenth-century French prints. Remarkably, the integrality of the text is distracted by a bizarre line adjustment. In reality, the twelfth line “è que la moitié des Etudes d'un Pein”

should have continued with line twenty two: “| tre , {uffit pour devenir Sculpteur. L'éten-”, but instead is followed by line thirteen “duë de la Peinture eft immentfe , la Sculpt=”, which actually should have succeeded line twenty two. For the spotting of words, such an error is hardly decisive, but it is for word co-occurrences. We could gain the best possible result with a third, improved version of Tesseract (ill. 3c).

While numbers can now be ‘read’ with more precision by the machine, and the lines almost exactly imitate the actual content of the page, other aspects – like the recognition of italics – came out worse again. All in all, one can nevertheless say that such a result permits the application of software solutions like AntConc or Voyant<sup>18</sup> in the intended scope. On the level of time alone, this positive outcome came at an expensive cost, as the process of enhancing, reading and evaluating meanwhile took approximately six minutes per page, which brought me to the following calculation: at that time, given a sum of 470 PDFs with an average of 300 pages, the working process for one PDF alone would take 1800 minutes or 30 hours. Of course, an average processing power of approximately 846.000 minutes or 14.100 hours was simply unfeasible in light of all the other projects maintained by our IT-team. Therefore, we had to decide, whether or not the project would have to be continued. What we were looking for was a return to a manageable number of ‘nows’.

### 3. Strategies to reduce and to expand the corpus

The guiding question was how the extensiveness of the text corpus could be reduced as much as possible. One part of the solution was as trivial as it was analogue. It turned out that for our purposes only a smaller part of the magazine’s text corpus was truly useful. Texts on art typically constitute only a very small percentage of journals like the *Mercure de France*<sup>19</sup> or the

---

18 <https://voyant-tools.org> [accessed: 24.04.2019].

19 An overview on the digitized versions of the *Mercure the France* between 1724 and 1778 can be found here: <http://gazetier-universel.gazettes18e.fr/periodique/mercure-de-france-1-1724-1778> [accessed: 27.05.2019]. The issues printed between 1778 and 1791 can be found here: <http://gazetier-universel.gazettes18e.fr/periodique/mercure-de-france-2-1778-1791> [accessed: 27.05.2019].

*Teutsche Merkur*.<sup>20</sup> We, therefore, decided to let a student assistant<sup>21</sup> preselect all art relevant passages in the magazines by usually starting from the table of contents. Of course, not every magazine still had its table of contents, which is why skimming the entire magazine was in any case obligatory. Nevertheless, we preferred to run the risk of overlooking texts to gather a relevant text quantity within a reliable period of investigation rather than being stuck with only a few digitized magazines that would have been the output of the limited resources we had for the whole project. Finally, we could bring together relevant material from the *Mercure the France*, printed between 1724 and 1756,<sup>22</sup> and from the *Teutsche Merkur*, published between 1773 and 1789. Admittedly, these periods of time do not illustrate art connoisseurship over a whole century, but they can, nonetheless, exemplify the prolific stages of connoisseurial practices in eighteenth-century France and Germany. Needless to say, the choice of the time spans also depends on capacity factors such as available digital resources and the workforce offered by the SFB.

Which other methods were available to reduce the corpus? It unquestionably appeared to be rather counter-intuitive to preselect passages in treatises on art and aesthetics. Although greater parts of, for example, Kant's *Critique of Judgement*<sup>23</sup> may have no relevance for the importance of practices of comparing in eighteenth-century connoisseurship, others could be all the more inspiring. A preselection here means to skip maybe the most interesting references. Then again, also in the field of significant art treatises it is possible to value more and less relevant material; to start the digitization with the most discussed ones proved to be a convenient approach.

A useful strategy not to reduce but to expand the corpus, however, is the concentration on eighteenth-century art treatises and/or magazines that have already been digitized referring to TEI-guidelines (Text Encod-

---

20 An overview on the digitized versions of the *Teutsche Merkur* between 1773 and 1789 can be found here: [http://ds.ub.uni-bielefeld.de/viewer/toc/1951387/1/LOG\\_0000/](http://ds.ub.uni-bielefeld.de/viewer/toc/1951387/1/LOG_0000/) [accessed: 17.04.2019].

21 I would like to thank our student assistant Felix Berge for his help.

22 The *Mercure the France* usually comprises six issues per year. In consequence, thirty years make a sum of almost two hundred issues with an average of four hundred pages per issue.

23 Kant, *Immanuel*, Critique of Judgement, ed. by Nicholas Walker, transl. by James Creed Meredith, Oxford [u. a.]: Oxford University Press, 1952.

ing Initiative).<sup>24</sup> They allow the user to analyze works in depth, for instance, in terms of a semantical structure, the recognition of a specific typology of comparisons,<sup>25</sup> and so forth. Unfortunately, only a small percentage of seventeenth- and eighteenth-century treatises and magazines with a focus on art and aesthetics available as scanned images or in PDF version is also accessible in fully digitized and linguistically marked .txt- or .xml-formats. For the research subject in question, databases like the *Observatoire de la vie littéraire*<sup>26</sup> or the *Deutsches Textarchiv*<sup>27</sup> offer a number of fully digitized texts that are central for eighteenth-century connoisseurship, as for instance Jean-Baptiste Dubos's *Réflexions critiques sur la poésie et la peinture*<sup>28</sup> or Johann Joachim Winckelmann's *Geschichte der Kunst des Alterthums*.<sup>29</sup> It may heuristically make sense to use such versions to exemplarily shed light on the reflections of practices of comparing itself, while the material digitized by means of dirty OCR could rather be used for quantitative questioning or to track down passages that could be of interest for the project.<sup>30</sup>

Of course, the more digitized, evaluated and corrected resources are at disposal, the more comprehensive the outcome will after all probability be. In case of the *Mercure the France*, for instance, the *Observatoire de la vie littéraire* offers a fully digitized and evaluated version of the forerunner magazine named *Mercure galant*.<sup>31</sup> This magazine was published between 1672 and 1710, and today 465 issues are available for a deeper analysis. Given the hypothesis that reflections on practices of comparing grew in the course of the eighteenth-century, one could expect an increase of certain key notions, as for instance “comparaison/comparer” or “jugement/juger”, or the co-occurrence of such notions. Therefore, it seems utterly useful to expand the corpus of the *Mercure de France* by the corpus of the *Mercure galant* to visualize the keyword's use over the years. Such a juxtaposition, however, brings

---

24 <https://tei-c.org/guidelines/> [accessed: 25.04.2019].

25 A good example for such an approach is Olga Sabelfeld's contribution in this volume.

26 <http://obvil.sorbonne-universite.site> [accessed: 23.05.2019].

27 <http://www.deutschestextarchiv.de> [accessed: 23.05.2019].

28 [https://obvil.sorbonne-universite.fr/corpus/critique/dubos\\_critiques](https://obvil.sorbonne-universite.fr/corpus/critique/dubos_critiques) [accessed: 23.05.2019].

29 [http://www.deutschestextarchiv.de/book/show/winckelmann\\_kunstgeschichte01\\_1764](http://www.deutschestextarchiv.de/book/show/winckelmann_kunstgeschichte01_1764) [accessed: 23.05.2019].

30 See for instance the projects by Christine Peters and Malte Lorenzen in this volume.

31 <https://obvil.sorbonne-universite.fr/corpus/mercure-galant/> [accessed: 23.05.2019].



its own challenges, because data repositories like the *Observatoire de la vie littéraire* are equipped with their own search function but cannot be extrapolated to download .txt-files (or other data formats). Although it is possible to obtain the data by means of the source code, such an indirect download process, once again, takes quite some time. Another problem lies in the comparability of the data itself because on the one hand, in addition to art critiques the downloaded data in the *Mercure galant* also comprises articles on poetry, theatre, history, politics and disputes discussed in the *Académie*. The repository of texts from the *Mercure the France* used in our project, on the other hand, includes only those parts on art critique. While it was necessary to reduce the corpus for processing dirty OCR, it now proves itself a disadvantage, because the much bigger corpus of the *Mercure galant* would have to be reduced accordingly. It is questionable, though, whether or not such a time exposure is worth the result.

This is why a tentatively carried out first comparison between the hits for the root word “compar” in the *Mercure galant* and the *Mercure the France* has only proceeded by its total numbers. Unfortunately, little can be concluded from the “concordance plot” for the regular expression search with the root word “compar”.<sup>32</sup> From the 465 issues of the *Mercure galant*, evidently, some issues have much more hits than others but a steady increase can not be established (ill. 4a, b).

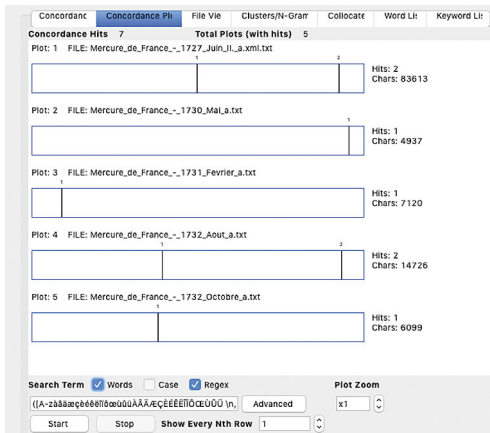
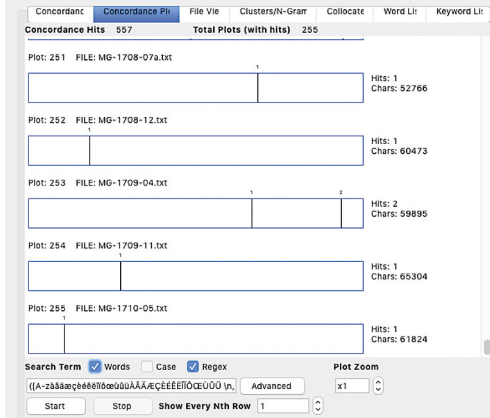
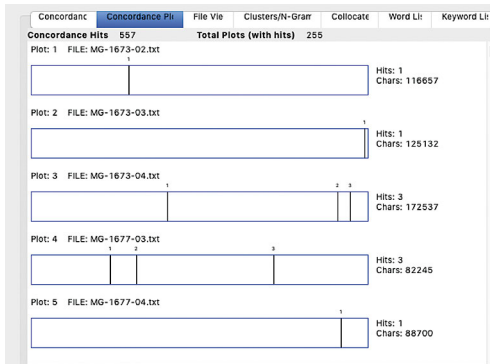
Evidently, the number of hits in the first years around 1673 is as high as around 1710. The same is true for the search in the extracted contributions on art critique in the *Mercure de France* (ill. 4c).

Here, only five out of thirty-six issues have a hit for the root word in question.<sup>33</sup> The situation in the *Mercure galant* is more or less similar given a hit rate of 255 out of 465 issues and considering that only a smaller part of the content is devoted to art critiques as such. Only one out of nine hits evident in the five plots of ill. 4a represents the root word “compar” in a text on art. All other hits are related to different topics such as theatre or music. At first

32 The regular expression used for the search in AntConc was provided by Stephan Porada. It is: “[A-zâãäæçèéêëïïòœùüÛÄÅĀÆÇĚĚĚĚİİÖÙÜÛ \n,;'\d(\)><-]\*compar[a-zâãäæçèéêëïïòœùüÛÄÅĀÆÇĚĚĚĚİİÖÙÜÛ\|n)\*compar[a-zâãäæçèéêëïïòœùüÛÄÅĀÆÇĚĚĚĚİİÖÙÜÛ]\* [A-zâãäæçèéêëïïòœùüÛÄÅĀÆÇĚĚĚĚİİÖÙÜÛ \n,;'\d(\)><-]\*”. The root word “compar” can be exchanged by any other root word.

33 When I tested the extracted text passages in AntConc, I could only refer to the first thirty-six out of altogether 110 articles on art published between June 1724 and October 1754.

III. 4a & 4b & 4c



sight, the result of the comparison appears to be disappointing but it is not given the astonishing outcome of a stability of hits in the course of time. This can either mean that the starting hypothesis on the increase of reflections on practices of comparing in the course of the eighteenth-century is probably wrong or that it cannot be shown by means of the use of words that indicate such a reflection. Another error source could perhaps be linked to the length of the examined periods, which might have been too short for substantial statements.

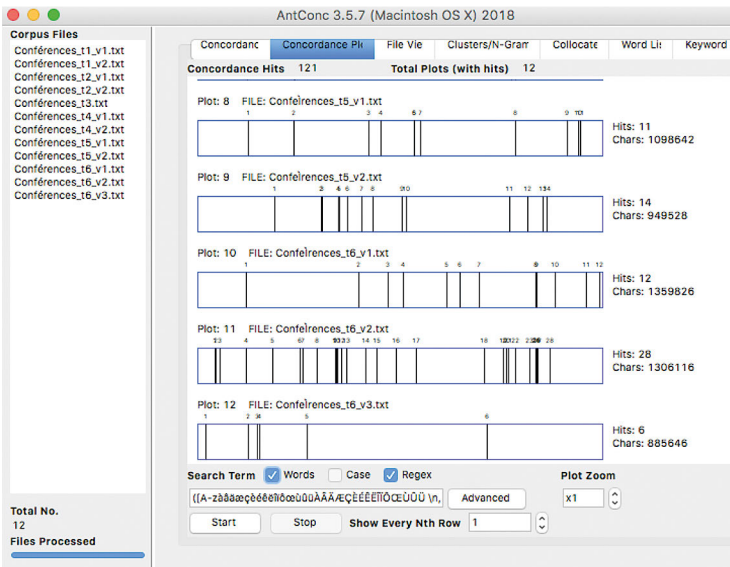
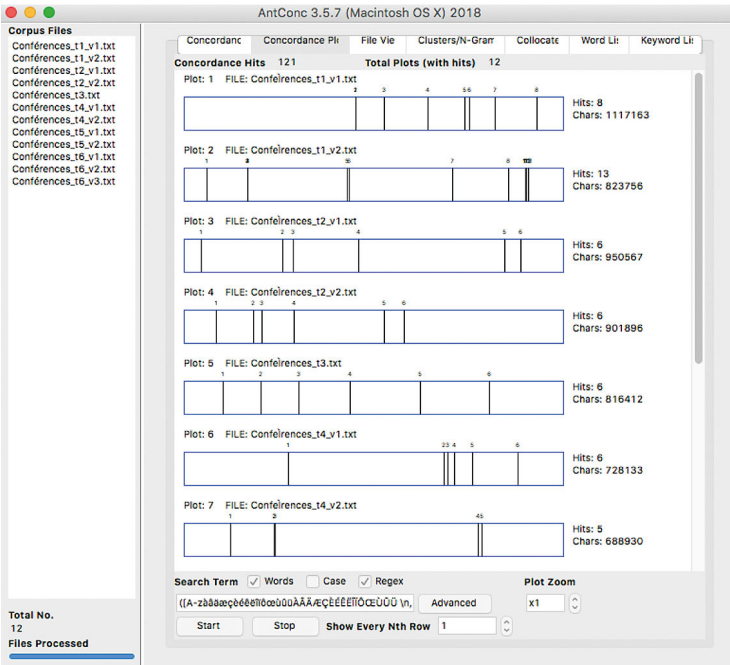
In other cases, it is possible to quantify the increasing use of notions, as can be seen in the following example. In the course of a contribution to the conference “Media of Exactitude” held in Basel in 2018,<sup>34</sup> I firstly asked for the occurrence of the word “exactitude” (and its derivatives) and, secondly, the co-occurrence<sup>35</sup> of the words “comparer/comparison” (and its derivatives) and “exactitude” (and its derivatives) in the *Conférences de l'Académie Royale de Peinture et de Sculpture* published between 1648 and 1792. My decision for choosing the lectures of the *Conférence* was due to the fact that they reflect in a way interests, fashions and focal points of well-known French connoisseurs for more than a century. Moreover, they are accessible as digitized, edited, and optical-character-recognized data.<sup>36</sup> Although the result is far from being representative, in the chosen span of time the word “exactitude” (ill. 5a, b) was increasingly used, as can be proved by comparing the hits in the early years (ill. 5a) with those in the later years (ill. 5b). The same is true for the co-occurrence of the word “comparison” (ill. 5c, d), which is optionally followed by the word “exactitude”. What is remarkable is the fact that in both cases only by the middle of the century an increase becomes visible.

34 For the conference, cf. <https://www.genauigkeit.ch/talk/conference/> [accessed: 12.05.2019].

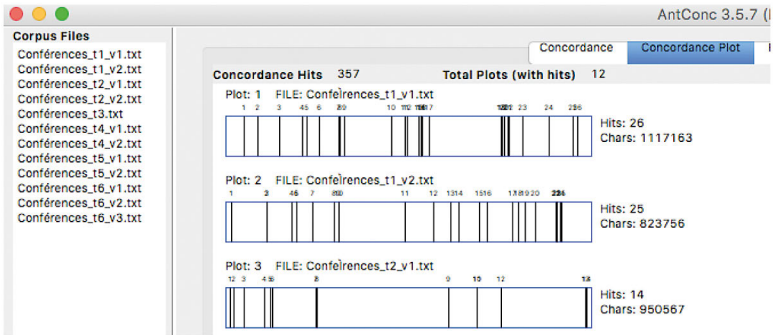
35 The regular expression used for the search in AntConc was provided by Stephan Porada. It is: “([A-zâãäæçèéëîïòœùüÿÄÅÆÇÈÉÊËÏÔÖÙÚ ò,;:’\d\(\)\><-]\*)|compar[a-zâãäæçèéëîïòœùüÿÄÅÆÇÈÉÊËÏÔÖÙÚ\|n]\*”)compar[a-zâãäæçèéëîïòœùüÿÄÅÆÇÈÉÊËÏÔÖÙÚ\|n]\*”)compar[a-zâãäæçèéëîïòœùüÿÄÅÆÇÈÉÊËÏÔÖÙÚ\|n]\*”) ([A-zâãäæçèéëîïòœùüÿÄÅÆÇÈÉÊËÏÔÖÙÚ ò,;:’\d\(\)\><-]\*)”(juge[af-zâãäæçèéëîïòœùüÿÄÅÆÇÈÉÊËÏÔÖÙÚ\|n]\*|[A-zâãäæçèéëîïòœùüÿÄÅÆÇÈÉÊËÏÔÖÙÚ ò,;:’\d\(\)\><-]\*)? [A-zâãäæçèéëîïòœùüÿÄÅÆÇÈÉÊËÏÔÖÙÚ ò,;:’\d\(\)\><-]\*)”. The root words “compar” and “juge” can be exchanged by any other root words.

36 The entire body of lectures given at the Parisian *Académie royale de peinture et de sculpture* were made available by an editing research project at the German Center for Art History (DFK Paris). For the fully digitized corpus, cf. <https://dfk-paris.org/en/research-project/editing-and-publishing-conférences-de-l'académie-royale-de-peinture-et-de> [accessed: 23.05.2019].

III. 5a & 5b



Ill. 5c & 5d



#### 4. How to handle first results?

First examinations of the isolated and OCR-processed passages in the *Teutsche Merkur* from 1773 to 1789 have shown that, for instance, the word “Vergleich/Vergleichung/vergleichen” can hardly be found. This holds particularly true because of the Fraktur, whose majuscule letters “B” and “V”, for example, look very similar. Therefore, using string characters like “rgleich” instead of root words like “Vergleich” or “Bergleich” undoubtedly yields more results. While doing so, I stumbled upon a text, which had only one hit for “Bergleichung”, but proved being a revealing example for practices of comparison in the close reading. The text “Ueber Christus und die zwölf Apostel, nach Raphael von Mark=Anton gestochen, und von Herrn Prof. Langer in Düsseldorf kopiert” was published in the fourth issue of the *Teutsche Merkur* in 1789.<sup>37</sup> As an apparent advertisement, the text of an anonymous author stresses the advantage of a series of engravings after Marcantonio Raimondi’s *Apostles* that shall help the beholder to refresh the vision of Raphael’s ingenious inventions.<sup>38</sup> The print series was copied by the artist Johann Peter von Langer (1756–1824) shortly before 1789. While the greater part of the review describes the prints and inventions themselves, in a subsequent passage the author discusses the value of the copies. They would inspire a fresh appraisal of the prints:

“These sheets arguably give us insight into the notion of the value of the originals in regard to invention, posture, drapery, character of hairs and faces. We can safely claim that no amateur of arts should fail to purchase these copies by Langer, even if he as an exception already possessed the originals [i. e., the prints by Marcantonio Raimondi]. In that case, the copies would still give some food for thought like a good translation.”<sup>39</sup>

---

37 *Anonymus*, Ueber Christus und die zwölf Apostel, nach Raphael von Mark=Anton gestochen, und von Herrn Prof. Langer in Düsseldorf kopiert, in: *Teutscher Merkur* 4 (1789), 269–277.

38 The connection between Raphael and Raimondi was demonstrated in-depth by: *Bloemacher, Anne*, Raffael und Raimondi, Berlin [et al.]: Deutscher Kunstverlag, 2016.

39 Translation by the author. The original quote is: “Diese Blätter gewähren also uns streitig einen Begriff von dem Werth der Originale in Absicht auf Erfindung, Stellung, Wurf der Falten, Charakter der Haare und der Gesichter, und wir dürfen wohl sagen, daß kein Liebhaber der Künste versäumen sollte, sich diese Langerischen Copien anzuschaffen,

At a time when the concept of originality evolved into the most important scheme for artistic achievements,<sup>40</sup> such praise for efficacy of the copy is rather astonishing. Moreover, the author underlines the fact that it can be enlightening to compare the original prints by Raimondi with the copies by Langer to expose still more the creators' artistic understanding and their – i. e., Raphael's and Raimondi's – light and fortunate nature.<sup>41</sup> The call for a visual comparison (“Vergleichung”) is followed by a meticulous analysis of the differences between Langer's copies and the original prints (ill. 6a, 6b); it brings out the tendency of connoisseurship to not only get the most complete possible overview but particularly to fragment the objects of research into small comparable entities. Langer's prints were used as a foil to foster the ideal-typical execution and planning of the original prints right into the folds and hatchings (“In den Originalen ist keine Falte, von der wir uns nicht Rechenschaft zu geben getrauen”). This and many other aspects make the short review a promising example for applied connoisseurship, in terms of the contemporary expectations of comparisons, of medial considerations as well as of the explicit guidance for a comparative approach.

The chance is rather modest that I would have found this review in the seemingly endless issues of the *Mercure de France* and the *Teutsche Merkur*. I could, however, track it down very quickly thanks to dirty OCR and the hit list in the AntConc concordance plot. Other than quantitative assertions, a qualitative perspective necessarily requests a thorough recontextualization of the extracted file. In the discussed case, the result could not be more surprising, because the anonymously published text is well known in the domain of research on Johann Wolfgang Goethe's art critical writings.<sup>42</sup> The review was

---

selbst in dem seltenen Falle wenn er die Originale besäße; denn auch alsdann würden ihm diese Copien, wie eine gute Uebersetzung, noch manchen Stoff zum Nachdenken geben.” cf. *Anonymus*, Ueber Christus und die zwölf Apostel, nach Raphael von Mark=Anton gestochen, und von Herrn Prof. Langer in Düsseldorf kopiert, 275.

40 Cf. for instance: *Mortier, Roland*, L'originalité: une nouvelle catégorie esthétique au siècle des lumières. Histoire des idées et critique littéraire, Genf: Droz, 1982.

41 The original quote is: “bey dem größten Kunstverstand, ein so leichtes und glückliches Naturell ihrer Urheber, daß sie uns wieder unschätzbar vorkommen.” cf. *Anonymus*, Ueber Christus und die zwölf Apostel, nach Raphael von Mark=Anton gestochen, und von Herrn Prof. Langer in Düsseldorf kopiert, 275.

42 Cf. *Osterkamp, Ernst*, Bedeutende Falten. Goethes Winckelmann-Rezeption am Beispiel seiner Beschreibung von Marcantonio Raimondis Apostelzyklus, in: Thomas W. Gaehtgens (ed.), Johann Joachim Winckelmann, 1717–1768, (Studien zum achtzehnten

Ill. 6a &amp; 6b



written by Goethe subsequent to his Italian Journey, and, apparently, the publisher of the *Teutsche Merkur*, Christian Martin Wieland, could temporarily provide him with the original prints by Marcantonio Raimondi.<sup>43</sup> It is stunning to see that the text was already discussed with regard to practices of comparative vision in Johannes Grave's seminal work on Goethe as a collector of prints. One could therefore say that such a positive result proves the efficacy of the approach to record relevant hits for a vocabulary reflecting comparisons as comprehensively as possible in different kinds of sources. However, it will hardly be possible to recontextualize every hit – or in other words every 'now' – towards a multiplicity of 'nows'. But it is, of course, possible to selectively densify certain "nows". Such a process would also comprise a reference to the visual resource itself (ill. 6a, 6b). Today, Langer's prints are surprisingly difficult to find within

---

Jahrhundert, 7), Hamburg: Meiner Verlag, 1986, 265–288; Osterkamp, Ernst, Im Buchstabenbilde: Studien zum Verfahren Goethescher Bildbeschreibungen (Germanistische Abhandlungen, 70), Stuttgart: Metzler, 1991, 54–71; Grave, Johannes, Der 'ideale Kunstkörper': Johann Wolfgang Goethe als Sammler von Druckgraphiken und Zeichnungen (Ästhetik um 1800, 4), Göttingen: Vandenhoeck + Ruprecht, 2006, 240–243.

43 J. Grave, Der 'ideale Kunstkörper', 240.



the image repositories provided by libraries, museums, and so forth. This circumstance evokes the lack of any illustration in publication formats like the *Teutsche Merkur* and it underlines the fact that Goethe's review actually might have prompted readers to make the purchase. It also shows how much it differs from what Peter Bell<sup>44</sup> has baptized 'digital connoisseurship': while the potentials of machine learning image recognition rest on a basis of hundreds of thousands of images online, in Goethe's time for the majority of people prints were still a rare commodity that had to be assembled in cumbersome collections. It could be said that at that time every comparison had its own value.

## Conclusion

The here proposed heuristic focuses not only on one specific method but seeks to combine different approaches, namely big data analysis and close readings, in reaction to both the diverse condition of the digitized material as well as the potentials offered by already fully digitized and OCR-processed open source material. One goal was to bring together as much material as possible to be able to trace changes in a long-term perspective with regard to the number of hits of vocabulary reflecting comparisons. At the same time, close reading enables recontextualization of the gathered bits and pieces and to switch from a quantitative to a qualitative argument. The application of a big data approach has been proven as a reliable 'good nose' for texts that could be crucial for a project on practices of comparison in connoisseurship. It turned out that dealing with the 'eternal nows' not only helped us to strategically pre-select the material but also to encourage entirely new questions. In the long term and also in light of a cost/benefit ratio it seems to have been worth the effort, because the established digital library allows even many more ways to question the once assembled material. It could be a next step to publish the library of OCR-processed texts on art and aesthetic and to share them with the scientific community. Such plans, however, have to be implemented in accordance with legal obligations, financial aftercare requirements, and so forth, so that one is reminded anew of the motto: "Tene mensuram".

---

44 Bell, Peter/Ommer, Björn, Digital Connoisseur? How Computer Vision Supports Art History, in: Stefan Albl/Alina Aggujaro (eds.), *Connoisseurship nel XXI secolo. Approcci, Limiti, Prospettive*, Rome: Artemide, 2016, 187–197.

## Bibliography

- Anonymus*, Ueber Christus und die zwölf Apostel, nach Raphael von Mark=Anton gestochen, und von Herrn Prof. Langer in Düsseldorf kopiert, in: Teutscher Merkur 4 (1789), 269–277.
- Bell, Peter/Ommer, Björn*, Digital Connoisseur? How Computer Vision Supports Art History, in: Stefan Albl/Alina Aggujaro (eds.), Connoisseurship nel XXI secolo. Approcci, Limiti, Prospettive, Rome: Artemide, 2016, 187–197.
- Bloemacher, Anne*, Raffael und Raimondi, Berlin [et al.]: Deutscher Kunstverlag, 2016.
- Borges, Jorge Luis*, The Library of Babel, in: Ficciones, transl. by Anthony Kerrigan et al., New York: Grove Press, 1962.
- Grave, Johannes*, Der 'ideale Kunstkörper': Johann Wolfgang Goethe als Sammler von Druckgraphiken und Zeichnungen (Ästhetik um 1800, 4), Göttingen: Vandenhoeck + Ruprecht, 2006.
- Kant, Immanuel*, Critique of Judgement, ed. by Nicholas Walker, transl. by James Creed Meredith, Oxford [u. a.]: Oxford University Press, 1952.
- Kohle, Hubertus*, Digitale Bildwissenschaft, Glückstadt: Verlag Werner Hülsbusch. Fachverlag für Medientechnik und -wirtschaft, 2013.
- Mortier, Roland*, L'originalité: une nouvelle catégorie esthétique au siècle des lumières. Histoire des idées et critique littéraire, Genf: Droz, 1982.
- Osterkamp, Ernst*, Im Buchstabenbilde: Studien zum Verfahren Goethescher Bildbeschreibungen (Germanistische Abhandlungen, 70), Stuttgart: Metzler, 1991.
- Osterkamp, Ernst*, Bedeutende Falten. Goethes Winckelmann-Rezeption am Beispiel seiner Beschreibung von Marcantonio Raimondis Apostelzyklus, in: Thomas W. Gaehtgens (ed.), Johann Joachim Winckelmann, 1717–1768, (Studien zum achtzehnten Jahrhundert, 7), Hamburg: Meiner Verlag, 1986, 265–288.
- Schwandt, Silke*, Looking for 'Time' and 'Change': Visualizing History in the Digital Age (draft version, forthcoming).
- Siegel, Steffen*, Tabula: Figuren der Ordnung um 1600, Berlin: Akademie-Verlag, 2009.

## Illustrations

- Ill. 1a: Page from the *Mercure de France*, Aoust 1729, 1785 © <http://gazetier-universel.gazettes18e.fr/periodique/mercure-de-france-1-1724-1778> [last access: 12.6.2019].
- Ill. 1b: .txt-file gained with a commercial OCR tool by 'reading' ill. 1a © Screenshot by the author
- Ill. 2a: Binarized page from the *Mercure de France*, Aoust 1729, 1785 © Screenshot by the author
- Ill. 2b: .txt-file gained with a commercial OCR tool by 'reading' ill. 2a © Screenshot by the author
- Ill. 3a: .txt-file gained with a first version of an improved Tesseract OCR tool © Screenshot by the author
- Ill. 3b: .txt-file gained with a second version of an improved Tesseract OCR tool © Screenshot by the author
- Ill. 3c: .txt-file gained with a third version of an improved Tesseract OCR tool © Screenshot by the author
- Ill. 4a: AntConc concordance plot of hits for the occurrence of the root word "compar" in the *Mercure galant* between 1673–1677 © Screenshot by the author
- Ill. 4b: AntConc concordance plot of hits for the occurrence of the root word "compar" in the *Mercure galant* between 1708–1710 © Screenshot by the author
- Ill. 4c: AntConc concordance plot of hits for the occurrence of the root word "compar" in the *Mercure de France* between 1727–1732 © Screenshot by the author
- Ill. 5a: AntConc concordance plot of hits for the occurrence of the root word "exactitude" in the *Conférence* between 1648–1746 © Screenshot by the author
- Ill. 5b: AntConc concordance plot of hits for the occurrence of the root word "exactitude" in the *Conférence* between 1747–1792 © Screenshot by the author
- Ill. 5c: AntConc concordance plot of hits for the co-occurrence of the root words "compare" and "exactitude" in the *Conférence* between 1648–1746 © Screenshot by the author

- Ill. 5d: AntConc concordance plot of hits for the co-occurrence of the root words “compare” and “exactitude” in the *Conférence* between 1747–1792  
© Screenshot by the author
- Ill. 6a: St Paul, Johannes Peter von Langer (after a print by Marcantonio Raimondi), etching, 216 × 135 mm (sheet), c. 1789, Wolfenbüttel, Herzog August Bibliothek, Graph.A1:1470 © <http://diglib.hab.de?grafik=graph-a1-1470> [last access: 12.6.2019].
- Ill. 6b: St Paul, Marcantonio Raimondi (after an invention by Raphael), etching, 215 × 141 mm (sheet), c. 1520, Wolfenbüttel, Herzog August Bibliothek, MRaimondi AB 3.26 © <http://kk.haum-bs.de/?id=raim-m-ab3-0026> [last access: 12.6.2019].



# Text Mining, Travel Writing, and the Semantics of the Global

## An AntConc Analysis of Alexander von Humboldt's *Reise in die Aequinoktial-Gegenden des Neuen Kontinents*

---

Christine Peters

Literary scholars have long been arguing that an integral part of travel writing is its continuous engagement with the world. In her introduction to travel writing, Anne Fuchs even identifies *Welthaltigkeit* – a containment of the world and an engagement with the world in the text – as one of the two key characteristics of travel writing, the other being its engagement with alterity.<sup>1</sup> While the interest in literary figurations of the global is not exclusive to research on travel writing,<sup>2</sup> the genre appears to be an especially productive object of investigation. Studies on Alexander von Humboldt, who can be considered a central figure of 19th century travel writing, frequently address the different ways in which he inscribes a global perspective into his writing, aiming to describe the world as a coherent, wholesome entity.<sup>3</sup> Even

---

1 Fuchs, Anne, Reiseliteratur, in: Dieter Lamping (ed.), *Handbuch der literarischen Gattungen*, Stuttgart: Alfred Kröner, 2009, 593–600.

2 Moser, Christian/Simonis, Linda, Einleitung: Das globale Imaginäre, in: Sebastian Moser/ Linda Simonis (eds.), *Figuren des Globalen: Weltbezug und Welterzeugung*, Göttingen: V&R unipress, 2014, 11–22.

3 Daum, Andreas, Alexander von Humboldt, die Natur als 'Kosmos' und die Suche nach Einheit: Zur Geschichte von Wissen und seiner Wirkung als Raumgeschichte, in: *Berichte zur Wissenschaftsgeschichte* 22 (2000), 246–250; Böhme, Hartmut, Ästhetische Wissenschaft: Aporien der Forschung im Werk Alexander von Humboldts, in: Ottmar Ette (ed.), *Alexander von Humboldt: Aufbruch in die Moderne*, Berlin: Akademie-Verlag, 2001, 17–32; Heyl, Bettina, *Das Ganze der Natur und die Differenzierung des Wissens: Alexander von Humboldt als Schriftsteller*, Berlin/Boston: De Gruyter, 2007, 179 ff.; Ette, Ottmar, *Humboldt und die Globalisierung: Das Mobile des Wissens*, Frankfurt a. M./Leipzig: Insel,

though these studies focus on different aspects of Humboldt's engagement with the world, methodically, they take the same traditional 'close-reading' approach, focusing on so called 'symptomatic readings' of the material. The following paper is conceived as a computationally-assisted contribution to thinking about the specific ways in which Humboldt imagines the world and globality in a broader sense. It focuses on one specific piece of travel writing, namely Herman Hauff's influential German translation of the *Relation Historique* (1814 ff.), the *Reise in die Äquinoktial-Gegenden des neuen Kontinents* (1859).<sup>4</sup> I proceed from the assumption that we can indeed gain some further insight into how Humboldt's travel writing engages with the world by identifying a set of relevant lexical features throughout a relatively small-sized corpus. Using *AntConc*<sup>5</sup> as a concordancing and text analysis toolkit, I argue for the interpretative productivity of combining 'distant' reading methods with a more traditional 'close' reading approach.

Firstly, the following paper employs computational methods to test a particular hypothesis derived from my earlier, non-digital research on world knowledge in Humboldt's travel writing. I previously adopted the traditional close reading approach, focusing mainly on specific text strategies that engage with the world in the text, such as practices of comparing and narrating.<sup>6</sup> These studies show that in his travel writing, Humboldt primarily

---

2009, 17 ff.; Knobloch, Eberhard, Alexander von Humboldts Weltbild, in: HiN: Internationale Zeitschrift für Humboldt-Studien X (2009): 36 f.; Görbert, Johannes, Die Vertextung der Welt: Forschungsreisen als Literatur bei Georg Forster, Alexander von Humboldt und Adelbert von Chamisso, Berlin: De Gruyter, 2014, 157; Erhart, Walter, Chamissos Weltreise und Humboldts Schatten, in: Julian Drews et al., Forster – Humboldt – Chamisso: Weltreisende im Spannungsfeld der Kulturen, Göttingen: V&R unipress, 2017, 13–34.

4 This study is embedded in a dissertation project on German travel writing, including German translations of originally French, English and Russian travelogues that were available to German speaking communities in the nineteenth century. The following paper can be considered as a trial run of a computational analysis of my material, focusing on a small exemplary corpus, which could easily be repeated and adjusted for the original travelogue or even a larger corpus including a variety of different travelogues or their translations. For an exposé of the dissertation project, see: <http://www.uni-bielefeld.de/sfb1288/personen/person.html?persId=80276541> [accessed: 31.08.2019].

5 Laurence Anthony (2018). *AntConc* (Version 3.5.7) [Computer Software]. Tokyo, Japan: Waseda University. Available from <http://www.laurenceanthony.net/software> [accessed: 31.08.2019].

6 Peters, Christine, Reisen und Vergleichen: Praktiken des Vergleichens in Alexander von Humboldt's *Reise in die Äquinoktial-Gegenden des Neuen Kontinents* und Adam Johann von

imagines the world as a relational concept by means of comparison: The text usually engages with the world by describing perceived geographical entities that ‘make up’ the world and by comparatively identifying or establishing relationships between these entities. A sense of the global thereby seems to be achieved through implicit means, through a comparative text structure rather than through addressing questions of the global on an explicit, lexical level. The following paper aims to test this particular thesis by employing computational methods that allow a ‘distant’ re-reading of the material.

Furthermore, this paper uses computational methods to generate new ways of thinking about how Humboldt’s travel writing engages with questions of globality. I aim to show that concordancing toolkits such as AntConc can be used to bring attention to the different levels on which a text can deal with the world as a concept.

## Corpus, approaches, preliminary decisions

Considering AntConc’s capacity to perform fast and accurate searches in small and mid-sized corpora, I chose a relatively small corpus of approximately 1250 pages, consisting of the four volumes of Hermann Hauff’s German translation of the *Relation historique*. I aim to show that with a mostly supervised analytical approach<sup>7</sup> you can identify a set of significant lexical

---

Krusensterns *Reise um die Welt*, in: IASL 42 (2017), 441–455; Peters, Christine, Historical Narrative versus Comparative Description? Genre and Knowledge in Alexander von Humboldt’s *Personal Narrative*, in: Martin Carrier/Carsten Reinhardt/Veronika Hofer (eds.), *Narratives and Comparisons: Adversaries or Supporters in Understanding Science?*, (manuscript in preparation). While the first paper focuses on practices of comparing in the German translation of the *Relation historique* as well as in the *Ansichten der Natur*, the second paper analyzes practices of comparing and narrating in the *Personal Narrative*, the English translation of the *Relation historique*.

- 7 In general, we can distinguish between two types of approaches in the field of text mining and corpus linguistics: On the one hand, we can take an ‘unsupervised’ or ‘corpus-driven’ approach that starts the analysis with direct observations of the corpus without bringing pre-existent hypotheses to the corpus. On the other hand, we can take a more ‘targeted’ or ‘corpus-based’ approach that analyzes the corpus in order to test pre-existing hypotheses (and maybe develop some new insights based on the results). Since this study tests very specific hypotheses about globality in Humboldt’s travel writing, most analytical steps belong to the latter category, taking a more ‘targeted’ or ‘supervised’ ap-



features even in a small corpus such as this. The analysis approached the corpus on four different levels:

Firstly, as a preliminary step, I performed a short unsupervised inquiry<sup>8</sup> using the word list and the n-grams tools to map out frequent lexical recurrences that might be relevant to the general semantic structure of the travelogue and to also gain a foundation for some first guesses on how ideas of the global might factor into the predominant lexical layout of the text. Secondly, in a significantly more extensive step, I performed a supervised corpus analysis, using a short preselected list of search words with a global resonance to map out the most frequent words and word clusters with the clusters tool. An additional goal was to identify the hermeneutic limits of this approach and the measures that need to be taken to further verify or specify the preliminary results. In a third step, I analyzed the 'distant' reading results in their specific contexts, using AntConc's file view tool for a 'close' re-reading of the results. As the following analysis shows, this combination of computational methods and a 'close' reading of the material was a crucial step in mapping out the different ways in which the text engages with the world. As a last step, I tested the particular hypotheses based on that process by comparing my corpus to a reference corpus, consisting of the five volumes of Alexander von Humboldt's *Kosmos*. The comparison proved that texts can engage with the world on different levels and to different degrees.<sup>9</sup>

---

proach. For the differentiation between a 'corpus-driven' and a 'corpus-based' approach see Anthony, Laurence, A critical look at software tools in corpus linguistics, in: Linguistic Research 30 (2013), 142. For an example of a paper that differentiates between a 'more targeted' and an 'unsupervised' approach see Erlin, Matt, Topic Modeling, Epistemology, and the English and German Novel, in: Journal of Cultural Analytics (2017). Other than Lawrence, Erlin does not reflect on how these terms are usually used in the field of corpus linguistics but rather just uses the terms 'unsupervised' and 'targeted' to describe what Lawrence calls a 'corpus-driven' and a 'corpus-based' approach.

8 A word list and n-grams analysis can be considered an unsupervised or corpus-driven analytical step insofar as neither tool requires a search word list or other devices that depend on pre-existing research. However, because my *interpretation* of the results relied heavily on my previous research, this analytical step appeared to be somewhat supervised after all, even if to a lesser extent than the following analytical steps.

9 Many of the text mining techniques used in this paper, such as frequency analysis, collocations, or n-gram clustering, have already been evaluated as valuable tools for the analysis and interpretation of data. For example, Dawn Archer demonstrates how such quantitative methods can enable the comparison of genres, the analysis of the author's ideological stance, or the attribution of authorship. See Archer, Dawn, Data Mining and

As mentioned above, I used a preselected list of search words with a global resonance for the supervised parts of the analysis. This list was built on the work of specialists in the field as well as on my own research on comparison and world knowledge in Humboldt's travel writing. The first search words I added to the list were "Neue Welt" ["New World"] and "Alte Welt" ["Old World"] as well as "neuer Kontinent" ["new continent"] and "alter Kontinent" ["old continent"] because my previous close reading analysis of Humboldt's travel writing had provided evidence that all of these terms are included in comparative practices that produce a sense of the global in the text.<sup>10</sup> For the present study, the aim was to use computational methods to go beyond symptomatic readings of the text and instead trace these terms and their lexical variations throughout the whole length of the travelogue and potentially generate some new insights into how these terms conceptualize the world on an explicit, lexical level.

To broaden the analytical spectrum, I added the search terms "Welt" ["world"] and "Erde" ["earth"] to the initial list, not only because they seemed like obvious choices but also because experts in the field had previously relied on them to explain how Humboldt conceptualizes the world in his writing.<sup>11</sup> Concerning the term "Welt", the goal was to analyze both the occurrence of the term in its singular form as well as the wider variety of compound nouns, such as "Weltbewusstsein" ["global consciousness", "world consciousness"], "Welthandel" ["world trade"], "Weltkulturen" ["world cultures"], "Weltbegriffe" ["world concepts"], "Weltwissenschaft" ["world science"], or "Weltethos" ["global ethics"], which, as Ottmar Ette shows, play a central role in Alexander von Humboldt's different world concepts.<sup>12</sup> Adding the term

---

Word Frequency Analysis, in: Gabriele Griffin/Matt Hayler (eds.), *Research Methods for Reading Digital Data in the Digital Humanities*, Edinburgh: Edinburgh University Press, 2016, 72–92.

10 C. Peters, *Reisen und Vergleichen*, 450–455.

11 See footnote 3.

12 O. Ette, *Humboldt und die Globalisierung*; Ette, Ottmar, *Weltbewusstsein: Alexander von Humboldt und das unvollendete Projekt einer anderen Moderne*, Weilerswist: Velbrück, 2002. None of these compound nouns are easily translated since they carry very specific philosophical and epistemological connotations. For this paper, I used Ette's translations of the terms. For the translations of "Weltbewusstsein" ["global consciousness", "world consciousness"] see Ette, Ottmar, *Unterwegs zum Weltbewusstsein: Alexander von Humboldts Wissenschaftsverständnis und die Entstehung einer ethisch fundierten Weltanschauung*, in: *HiN: Internationale Zeitschrift für Humboldt-Studien* 1 (2000);

“Erde” to the list, including compound nouns in the same manner as with the term “Welt”, proved essential to the study since it allowed a focus on two alternative figurations of the global.<sup>13</sup> In both cases I decided not to add the specific compound nouns to the list of search words but rather their lexical stems “erd\*” and “\*welt\*”<sup>14</sup> to enable search results that include compound nouns that I had not anticipated. The final list<sup>15</sup> included all the search terms in form of their lexical stems:

---

*Ette, Ottmar*, Languages about Languages: Two Brothers and one Humboldtian Science, in: *HiN: Internationale Zeitschrift für Humboldt-Studien* XIX (2018), 47–61. For the translations of “Welthandel” [“world trade”] and “Weltkulturen” [“world cultures”] see *Ette, Ottmar*, The Scientist as Weltbürger: Alexander von Humboldt and the Beginning of Cosmopolitics, in: *HiN: Internationale Zeitschrift für Humboldt-Studien* II (2001). For the translation of “Weltbegriffe” [“world concepts”] see *Ette, Ottmar*, Unterwegs zu einer Weltwissenschaft? Alexander von Humboldts Weltbegriffe und die transarealen Studien, in: *HiN: Internationale Zeitschrift für Humboldt-Studien* VII (2006), 34–54. For the translation of “Weltwissenschaft” [“world science”] see *O. Ette*, Languages about Languages. For the translation of “Weltethos” [“global ethics”] see *O. Ette*, Unterwegs zum Weltbewußtsein.

- 13 *Stockhammer, Robert*, Welt oder Erde? Zwei Figuren des Globalen, in: Christian Moser/Linda Simonis (eds.), *Figuren des Globalen. Weltbezug und Welterzeugung in Literatur, Kunst und Medien*, Göttingen: V&R unipress, 2014, 47–72.
- 14 As most search engines, AntConc allows the truncation of search words, meaning the use of “wildcards”, such as \*, #, ?, or +. The asterisk (which finds zero or more characters) helps finding different variations of the same word, such as nouns in their singular and plural forms (e.g., “Welt” and “Welten” for “welt\*”) or different compound nouns containing the same words (e.g., “Erdball” and “Erdstrich” for “erd\*”). However, it is important to note that even though it is possible to search for different variations of the same word, the list of results will not display these variations as different instances of the same word. In addition, the list will most likely include other random words that contain the same combination of characters but do not belong to the same lexical group. In any case, it is necessary to sort through the list of results manually to identify the relevant findings.
- 15 I excluded the search term “\*erd\*” after my first attempts because it rendered too many results, namely 2000 occurrences in the corpus, which in turn included too many results not connected to the term “Erde”, such as “werden” or “allerdings”. I also excluded the search term “\*asi\*” for the same reason: Even though it only occurs 226 times in the corpus, too many instances were not related to the term “Asien”, such as the term “Brasilien” which occurred twice as often as the term “Asien”. Also, since the German differentiation between the upper and lower case was of no significance to my analysis, in the following, I quote all my search words and results in the lower case form, the only exception being quotes including whole text passages or longer phrases.

Fig. 1: Search word list for the distant reading analysis, lexical stems

*welt*	[world]
erde*, erd*	[earth]
neue* welt*, alte* welt*	[new world, old world]
neue* kontinent*, alte* kontinent*	[new continent, old continent]

## Distant reading I: unsupervised corpus analysis

Before approaching the corpus with the preselected search word list, I conducted an unsupervised corpus analysis. I used the n-grams tool which allows the user to find common expressions in a corpus by searching the entire corpus for “N” length clusters (e. g., one word, two words, etc.). For this corpus, the analysis rendered the most significant results at an n-gram size of four words:

Table 1: Results of the n-gram analysis, n-gram size of 4 words

AntConc 3.5.7 (Windows) 2018

File Global Settings Tool Preferences Help

Corpus Files

humboldt\_aequinoktia  
humboldt\_aequinoktia  
humboldt\_aequinoktia  
humboldt\_aequinoktia

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Total No. of N-Gram Types 413570 Total No. of N-Gram Tokens 427102

Rank	Freq	Range	N-gram
1	76	4	a v humboldt reife
2	64	4	in der heißen zone
3	54	4	an den ufern des
4	44	4	in der nähe der
5	44	4	von oft nach welt
6	42	4	von zeit zu zeit
7	41	4	auf dem rücken der
8	40	4	in der neuen welt
9	39	4	in der nähe des
10	37	4	an der mündung des
11	37	4	an ort und stelle
12	36	4	an der küfte von
13	32	4	aus der familie der

Search Term  Words  Case  Regex  N-Grams  Advanced

N-Gram Size Min. 4 Max. 4

Min. Freq. Min. Range 1 1

Sort by  Invert Order Search Term Position  On Left  On Right

Sort by Freq

Total No. 4

Files Processed

Clone Results

Apart from the highest ranking n-gram “a v humboldt reise”,<sup>16</sup> which refers to the title of the travelogue, and the temporal n-gram “von zeit zu zeit” [from time to time], all the n-grams ranking in the top ten carry spatial meaning, either referring to a specific region or place, or describing a spatial movement in a certain direction. The only n-gram with a global ring to it is the 8th ranking n-gram, “in der neuen welt” [in the new world], which, as mentioned before, evokes a sense of the global but does in fact refer to a specific region of the world and more specifically to the region that is the object of this travelogue.

Since continuously narrating movement through space is an integral part of travel writing, it is almost to be expected that the lexical features reflect this spatial focus of the text. However, for my analysis of the imagination of the global, it is noteworthy that none of the high ranking n-grams actually feature a global perspective on a lexical level. A general wordlist points to a similar focus on local or regional space. The wordlist tool counts all the words in a corpus and then displays them in an ordered list which gives the user easy access to the most frequent words in a corpus. With 1107 instances, the highest ranking word in the travelogue is “orinoko” [Orinoco], the river that poses a major stop on Humboldt’s travels and a central object of description in the travelogue. In comparison, the term “welt” only ranks 76th with 219 occurrences. While 219 occurrences can still be considered significant in a 1250 page corpus, the difference in frequency is still remarkable. It points to the idea that on the lexical surface of the text, globality does not actually play as big a role as it does when it comes to the overall narrative and comparative structure of the text. One way to test this tentative notion is to apply a supervised approach and analyze the terms “welt” and “erde” in the concordance, clusters, and collocates tools.

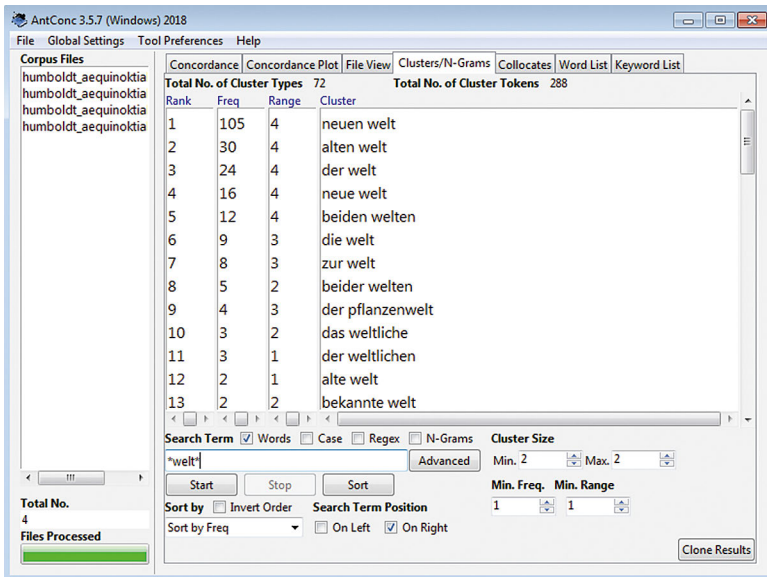
---

16 Literally this n-gram translates as “a[lexander] v[on] humboldt voyage”, but with regard to the title of the English translation, a more appropriate translation would be “personal narrative.” See *Humboldt, Alexander von, Personal Narrative of Travels to the Equinoctial Regions of the New Continent, during the Years 1799–1804*, 7 vols., translated by Helen Maria Williams, New York: Cambridge University Press, 2011, 1814 ff.

## Distant reading II: supervised corpus analysis

The supervised corpus analysis, at first, seemed to contradict the notion that the lexical features of the text do not point to the world in its entirety. After all, a search for “\*welt\*” in the corpus<sup>17</sup> shows 288 instances of the word.<sup>18</sup> However, a cluster analysis shows that the term carries mostly regional connotations. The clusters tool displays clusters containing a chosen search word, based on specific search conditions, such as the size of the cluster or the position of the search word. A cluster analysis for the search term “\*welt\*” with the search term positioned on the right and a cluster size of two words rendered the following results:

*Table 2: Results of the cluster analysis for “\*welt\*”, cluster size of two words, search word on the right*



Rank	Freq	Range	Cluster
1	105	4	neuen welt
2	30	4	alten welt
3	24	4	der welt
4	16	4	neue welt
5	12	4	beiden welten
6	9	3	die welt
7	8	3	zur welt
8	5	2	beider welten
9	4	3	der pflanzenwelt
10	3	2	das weltliche
11	3	1	der weltlichen
12	2	1	alte welt
13	2	2	bekannte welt

17 To determine how often a specific search word occurs in the corpus, I used the concordance tool which displays the search words in their contexts but also provides the user with the total number of occurrences.

18 The earlier mentioned 219 occurrences as displayed in the wordlist refer to the actual word form “welt”, while the above mentioned 288 occurrences refer to all lexical variations of the term as displayed in the concordance tool when searching for the term “\*welt\*”.

The list shows that different variations of the cluster “neue welt” [new world] as well as the cluster “alten welt” [old world] indeed figure prominently in the top 5 of the 72 existing clusters. According to a search in the concordance tool, the most frequent cluster, “neue\* welt\*”, occurs 121 times, while the second leading cluster, “alte\* welt\*”, occurs 32 times. This shows that on a lexical level “welt” is rarely used as a global but mostly as a regional and relational term that divides the planet into two corresponding worlds. The third ranking cluster, “beide\* welten” [both worlds], which occurs 18 times in the corpus, only reinforces this hemispheric notion of “welt”. With only 9 and 24 occurrences, the clusters “die welt” [the world] and “der welt” [declined form of “die welt”], which actually refer to the world in a generic and singular sense, occur significantly less frequent than the hemispheric clusters.

Changing the cluster size to one word shows that there are a number of compound nouns that refer to the world as a whole but that occur far less frequently than the above mentioned hemispheric clusters. A search for these compound nouns in the concordance tool, using search words such as “weltmeer\*” [ocean, world sea], “weltbeschreibung\*” [history or description of the world], “welthandel\*” [world trade], or “weltraum\*” [space, cosmos], shows that the majority of them occur less than five times across the entire length of the corpus. This supports the observation that the term “welt” appears to be mostly used as a hemispheric, regional concept rather than a global one. A comparison with another cluster analysis further corroborates this thesis. The term “kontinent\*”, an explicitly regional term that appears 91 times in the corpus, mostly carries the same attributes as the term “welt”: The most frequent clusters are “neue\* kontinent” [new continent], occurring 44 times, and “beide\* kontinente” [both continents], occurring 17 times.

Concerning the term “erde”, the supervised analysis pointed into a slightly different direction. While the above mentioned cluster analysis for “welt” mostly cycled back to terms I had already preselected for my search term list, the cluster analysis for “erd\*” supplied me with a number of terms that I had not anticipated:

Table 3: Results of the cluster analysis for “erd\*”, cluster size of one word, search word on the right

Rank	Freq	Range	Cluster
1	125	4	erde
2	107	4	erdbeben
3	37	4	erdstrich
4	33	4	erd
5	30	4	erdoberfläche
6	24	2	erdstöÙe
7	15	3	erdballes
8	12	4	erdfriche
9	11	4	erden
10	11	4	erdreich
11	10	2	erdstoÙ
12	9	3	erdbebens
13	9	4	erdfrichen
14	8	4	erdball
15	7	4	erdigen
16	7	3	erdfriches
17	6	2	erderfchütterungen
18	5	3	erdballs
19	4	3	erdboden
20	4	2	erdfläche

This list of the 20 highest ranking one-word clusters contains a number of reoccurring terms in their different grammatical variations, most prominently “erde” [earth, soil, ground], “erdbeben” [earthquake/s], “erdstrich” [zone, part of the earth], “erdball” [globe], “erdoberfläche” [earth’s surface], and “erdstöÙ” [earth tremor], whose exact frequency can be determined by means of the concordance tool: “erde” (125), “erdbeben\*” (116), “erdstrich\*” (69), “erdball\*” (30), “erdoberfläche\*” (30), “erdst\*Ù\*”<sup>19</sup> (39). I used this list as a starting point for my analysis which provided me with some unpredicted

19 Because the vowel in the word “erdstöÙ” changes into an umlaut in the plural (“erdstöÙe”) an asterisk had to be used in the middle of this search word (“erdst\*Ù\*”). This way, both the singular and the plural appear in the results.



and yet very productive analytical trajectories but also pointed to some of the hermeneutic limits of a computational approach.

At first glance, there appears to be a higher number of terms that might explicitly refer to the planet as a whole, such as “erde” [earth, soil, ground], which occurs 125 times in the corpus, or “erdball\*” [globe], which occurs 30 times. A second cluster analysis focusing on these two terms corroborated this observation since they mostly appear in a generic form that is quite different from the dualistic clusters “neue welt” [new world] and “alte welt” [old world]. The two highest ranking two-word clusters with the search word “erde” positioned on the right are “die erde” [the earth] with 37 hits and “der erde” [declined form of “die erde”] with 31 hits. In a similar manner, all 30 two-word clusters with the search term “erdball\*” [globe] on the right refer to the term in a generic and singular way with a frequency ranging between 14 and one occurrences, such as the differently declined forms of “der erdball” [the globe] (“des erdballes”, “dem erdball”, “den erdball”, “des erdballs”, and “dem erdballe”) or the differently declined forms of “unser erdball” [our globe] (“unseres erdballs” and “unseres erdballes”). However, since the term “erde” carries two meanings, referring to the planet as a whole as well as to the ground or a specific type of soil, some further testing was necessary. The cluster and collocates analyses did not help to determine how many of the 125 hits actually refer to the planet and how many to the ground. As the next chapter shows, in this case only an arduous use of the file view provided specific numbers, which ultimately contradicted the notion that the term “erde” refers to the whole planet more often than the term “welt”.

The initial list of clusters displayed in table 3 also contained the high ranking compound noun “erdstrich” [zone, part of the earth] which in all its grammatical variations occurs 69 times in the corpus. Semantically it refers to the earth in its totality but also points to a specific region inside this totality. A further cluster analysis showed that the term mainly refers to specific regions with regard to their climatic conditions, the most frequent two-word clusters being “heiße\* erdstrich\*” [torrid zone] with 31 occurrences and “gemäßigte\* erdstrich\*” [temperate zone] with 11 occurrences. The compound noun “erdstrich” does not only stress the regional particularity of the planet but more specifically its climatic diversity.<sup>20</sup>

---

20 In fact, the English translation frequently refers to what is called “erdstrich” in German both as “zone” and “climate.”

Two other words displayed in the initial list of clusters, “*erdbeben*” [earthquake/s] and “*erdstoß*”/“*erdstöße*” [earth tremor/s], refer to a geological discourse on volcanic activity. With 116 and 39 occurrences they feature prominently in the corpus, but whether they refer to volcanic activity on a global scale or to singular, regional earthquakes cannot be determined by means of distant reading. Research on the matter shows that Humboldt’s geological analysis of volcanic activity has a strong global focal point,<sup>21</sup> but without switching to close reading in the file view there is no way to decide whether this global focal point manifests on a lexical level. Whether these terms are used in a local or global way can only be determined by analyzing them in their specific contexts. The only tentative notion that is to be drawn from the cluster analysis is that on a lexical level the term “*erde*” has a strong geological connotation that the term “*welt*” lacks.

In sum, the supervised cluster analysis suggested that the term “*erde*” in the form of its different compound nouns points to the planet both in its totality and its particularity, unlike the term “*welt*” which mostly points to local particularity. The results also indicated specific geological and climatic connotations of the term “*erde*” which the term “*welt*”, at least on an explicit lexical level, does not carry.

In addition to the cluster analysis, I performed a co-occurrence analysis using the collocates tool, which allows the user to investigate non-sequential patterns in a corpus. The tool displays the collocations of a search word, meaning words that occur in the context of the given search word. The aim was to identify frequent co-occurring words and to see whether they confirmed the mostly regional connotations of the search terms or rather contradicted them. However, even when applying the maximum search window size of 40 words, the analysis rendered no viable results. None of the high ranking collocations allowed valid guesses on how the specific contexts interacted with the regional focus of the search words. Again, using the file view for a close reading analysis of the results appeared to be the necessary next step.

Taking previous research on Humboldt’s engagement with the world into account, the results of the supervised corpus analysis seem unremarkable

---

21 *Kraft, Tobias*, *Erdwissen im Angesicht der Berge: Die Vulkanlandschaften der Jorullo-Ebene als Heuristik der Geologie*, in: Ottmar Ette/Julian Drews (eds.), *Horizonte der Humboldt-Forschung: Natur, Kultur, Schreiben*, Hildesheim: Georg Olms, 2016, 97–124.

and remarkable at the same time. As Ottmar Ette shows, Humboldt's world concepts are based on a long tradition of perceiving the world as a divided, hemispheric space that actually consists of *two* worlds, the "new" and the "old" one.<sup>22</sup> The fact that the lexical level of the travelogue reflects this tradition is not surprising. However, as mentioned in the introduction, scholars have also frequently pointed towards the global perspective of Humboldt's research and writing, towards his aim to describe the world as a coherent, wholesome entity. What is remarkable about the results presented above is that the travelogue's lexical level does not display such a strong focus on the world in its totality, but rather leans strongly towards a territorial usage of the terms "welt" and "erde", drawing attention to specific regions of the planet rather than to the planet as a whole.

### **A close re-reading of the results: approaching the global on a different level**

In reaction to the distant reading results, the first purpose of the close reading approach was to further investigate the actual connotations of the search terms in their specific contexts. A second purpose was to look for patterns in these contexts, for potential text structures that go beyond the lexical layout of the text, and to see how the search terms were embedded in these text structures. The initial idea was to first approach such contextual matters by using the collocates tool. However, the tool's capacity appeared too limited for such an approach in two ways. First, the maximum search window size of 40 words proved to be too small, considering that an argument or a narrative sequence often develops over a whole paragraph or even over several pages. Second, a collocation analysis only takes re-occurrences of identical words into account, thereby only focusing on *lexical* context. Therefore, a collocation analysis does not necessarily detect recurring context in the form of *text strategies* that often vary when it comes to the specific words which are used. A close reading of the material therefore proved the necessary next step.

---

22 O. Ette, *Humboldt und die Globalisierung*, 36 ff.

As a first step, I preselected a new search word list based on the distant reading results, containing the most frequent words and word combinations from the cluster analysis:<sup>23</sup>

*Fig. 2: Search word list for the close reading analysis, lexical stems*

neue* welt	[new world]
beide* welten	[both worlds]
alte* welt	[old world]
erde	[earth]
erdball*	[globe]
erdstrich*	[zone, part of the earth]

I searched the corpus for these words using the file view tool which displays the search results in the individual text files and thereby allows the user to analyze them in their specific contexts. To keep my results comparable and also to keep the analysis practicable I decided that for each search hit I would consider the whole surrounding paragraph as context. Methodologically, it should also be noted that the patterns I identified relied heavily not only on this pragmatic decision but also on the specific questions I brought to the material. As mentioned in the introduction, my previous research had indicated that comparative text practices are especially productive when it comes to dealing with the world in its entirety.<sup>24</sup> One of my explicit goals was to test this thesis and see whether I could trace those comparative practices by analyzing the search terms in their contexts. With those questions in mind, I paid special attention to the spatial structure of the paragraphs

23 I did not include the search term “erd\*” in the list because it occurs 575 times in the corpus, making a thorough and continuous close reading analysis quite difficult. Instead, I decided to use frequent words and word clusters indicated in table 3 (“erde”, “erdball\*”, and “erdstrich\*”) as a starting point for my close reading analysis. That way the list of results was separated into significantly smaller chunks, which made a “manual” analysis much more effective. Furthermore, I did not add the geological words referring to earthquakes and seismic activity to the list, but rather focused on words that carried spatial connotations (again: “erde”, “erdball\*”, and “erdstrich\*”).

24 C. Peters, *Reisen und Vergleichen*; C. Peters, *Historical Narrative versus Comparative Description*.

and to whether – and if so how – they establish relationships between different geographical entities. Also, the numbers I provide relied on analytical decisions. They were not supplied by the tool itself but rather by my ‘manual’ interpretation of the results. In short, the findings of my close reading approach, even if carried out in the AntConc toolkit, are not the direct result of a computational analysis but rather of the hermeneutic attempt to organize the material in accordance to the above mentioned analytical interests.

Tracing the high-ranking cluster “neue welt” [new world] in the file view showed that at least 69 of the 121 occurrences are set in a global context. The biggest portion of these 69 occurrences, at least 50 according to my analysis, explicitly or implicitly shows a comparative context, as can be seen in the following exemplary excerpts:<sup>25</sup>

“Matanza bedeutet Schlachtbank, Blutbad, und schon das Wort deutet an, um welchen Preis der Sieg erkaufte worden. In der **Neuen Welt** weist er gewöhnlich auf eine Niederlage der Eingeborenen hin; auf **Tenerifa** bezeichnet das Wort Matanza den Ort, wo die Spanier von denselben Guanchen geschlagen wurden, die man bald darauf auf den spanischen Märkten als Sklaven verkaufte.”

[“Matanza signifies butchery, or carnage; and the word alone recalls the price, at which victory has been purchased. In the **New World**, it generally indicates the defeat of the natives; at **Teneriffe**, the village of Matanza was built in a place where the Spaniards were conquered by those same Guanches, who soon after were sold as slaves in the markets of Europe.”]<sup>26</sup>

“Ich bin weit entfernt, die Sprachen der **Neuen Welt** den schönsten Sprachen **Asiens** und **Europas** gleichstellen zu wollen; aber keine von diesen hat ein klareres, regelmäßigeres und einfacheres Zahlssystem als das Qquichua und

---

25 Typographically, the original text uses “f” for “s” in certain cases. Since typographical issues are of no significance to the present study, I adjusted the quotes, always using “s” even though AntConc displays the original characters. Furthermore, I highlighted the names of places and regions that are being compared or related to each other in order to enable the reader to follow the analysis more easily.

26 *Humboldt, Alexander von, Reise in die Aequinoktial-Gegenden des neuen Kontinents, vol. 1., translated by Hermann Hauff, Stuttgart: Cotta, 1859, 69. [A. v. Humboldt, Personal Narrative, vol. 1, 134.]*

das Aztekische, die in den großen Reichen Cuzco und Anahuac gesprochen wurden.”

[“I am far from placing the languages of the **New World** in the same rank with the finest languages of **Asia** and **Europe**; but no one of them has a neater, more regular, and simpler system of numeration, than the Qquichua and the Azteck, which were spoken in the great empires of Couzco and Anahuac.”]<sup>27</sup>

“Nach allem, was wir vom Gleichgewicht der Meere wissen, kann ich nicht glauben, daß die **Neue Welt** später als die **Alte** dem Schoße des Wassers entstieg, daß das organische Leben in ihr jünger, frischer sein sollte; wenn man aber auch keine Gegensätze zwischen den **zwei Halbkugeln** desselben Planeten gelten läßt, so begreift sich doch, daß auf derjenigen, welche die größte Wasserfülle hat, die verschiedenen Flußsysteme längere Zeit gebraucht haben, sich voneinander zu scheiden, sich gegenseitig völlig unabhängig zu machen.”

[“From what we know of the equilibrium of the seas, I cannot think, that the **New World** issued from the waters later than the **Old**; and that organic life is there younger, or more recent: but, without admitting oppositions between the **two hemispheres** of the same planet, we may conceive, that in the hemisphere most abundant in waters the different systems of rivers required more time, to separate themselves from one another, and establish their complete independence.”]<sup>28</sup>

“In der **Neuen Welt** gingen ähnliche Wanderungen in der Richtung von Nord nach Süd. In **beiden Halbkugeln** richtete sich die Bewegung der Völker nach dem Zug der Gebirge; aber im **heißen Erdstrich** wurden die gemäßigten Hochebenen der Kordilleren von bedeutenderem Einflusse auf die Geschichte des Menschengeschlechtes als die Gebirge in **Centralasien** und **Europa**.”

[“In the **New World** similar migrations flowed from north to south. Among the nations that inhabited the **two hemispheres**, the direction of this movement followed that of the mountains; but, in the **torrid zone**, the temperate

27 A. v. Humboldt, *Reise*, vol. 2, 22. [A. v. Humboldt, *Personal Narrative*, vol. 3, 242.]

28 A. v. Humboldt, *Reise*, vol. 3, 256. [A. v. Humboldt, *Personal Narrative*, vol. 5, 315.]

table-lands of the Cordilleras exerted a greater influence on the destiny of mankind, than the mountains of **Asia** and **central Europe**.”<sup>29</sup>

Throughout the corpus, text passages such as these compare the “neue welt” [new world] to other regions or places of the world, such as the “Alte Welt” [“Old World”], “Asien” [“Asia”], or “Europa” [“Europe”]. They do so either by describing them in a similar manner, thereby urging the reader to perform the actual comparison themselves, or by explicitly arranging them in a comparative manner, using adjectives in their comparative or superlative form, and naming specific similarities and differences.<sup>30</sup>

The rest of the 69 occurrences in a global context also relate the term “neue welt” [new world] to other regions of the world. As the following list of excerpts shows, they do not necessarily compare these regions to each other but rather focus on different types of relations between them:

“Erzeugnisse der **Neuen Welt** können in die **Alte Welt** nur in hohen Breiten und in der Richtung des **Stromes von Florida** gelangen.”

[“The productions of the **new world** cannot reach the **old**, but by the very high latitudes, and in following the direction of the **current of Florida**.”]<sup>31</sup>

“Wie das Zuckerrohr zuerst von den **Kanarien** in die **Neue Welt** kam, so stehen noch jetzt meist Kanarier oder Isleños den großen Pflanzungen vor und geben beim Anbau und beim Raffinieren die Anleitung.”

[“If the first canes arrived in the **New World** from the **Canary islands**, it is also in general Canarians, or Islenos, who are now placed at the head of

29 A. v. Humboldt, *Reise*, vol. 4, 234. [A. v. Humboldt, *Personal Narrative*, vol. 6, 15.]

30 Some of the examples that are not displayed here are much shorter, others too long to quote them in a paper as the one at hand. For example, Humboldt's extensive comparison of different plains across the globe stretches over a number of pages, in one paragraph alone referring to “Europa” [Europe], “Asien” [Asia], “Afrika” [Africa], “Amerika” [America], “Arabien” [Arabia], “Venezuela” [Venezuela], “Peru” [Peru], “Neuen Welt” [New World], and the specific plains and deserts “Dsungarei” [Dzungaria], “Pußten” [Puszta], “Gobi” [Gobi], “Sahara” [Sahara], and “Llanos” [Llanos]. See A. v. Humboldt, *Reise*, vol. 2, 269. [A. v. Humboldt, *Personal Narrative*, vol. 4, 295 f.]

31 A. v. Humboldt, *Reise*, vol. 1, 29. [A. v. Humboldt, *Personal Narrative*, vol. 1, 58.]

the great plantations, and who superintend the labours of cultivation and refining.”]<sup>32</sup>

“... große Naturforscher (Cuvier) scheinen anzunehmen, daß alle Pythone der **Alten**, alle Boa der **Neuen Welt** angehören.”

[“... great naturalists † appear to admit, that all the pythons belong to the **ancient**, and all the boas to the **New World**.”]<sup>33</sup>

“Gegenwärtig teilen sich, kann man wohl sagen, drei **Völker europäischer Abkunft** in das Festland der **Neuen Welt**: das eine, das mächtigste, ist germanischen Stammes, die beiden anderen gehören nach Sprache, Litteratur und Sitten dem lateinischen Europa an.”

[“The continental part of the **New World** is at present in some sort divided between three **nations of European origin**; one, the most powerful, is of Germanic race; the two others belong by their language, their literature, and their manners, to Latin Europe.”]<sup>34</sup>

As the list shows, a sense of globality is achieved in various ways, for example by describing global trade routes, by tracing the migration of agricultural crops, by debating to which region of the world certain animal species belong, or by discussing the European colonization of the Americas. These results indicate that the corpus actually shows a very strong focus on the world as a whole, but that this focus does not manifest on the level of lexical features. As the cluster analysis has shown, the corpus does not contain many lexical entities that refer to the world in its totality, whereas in the file view we can trace a number of comparative and otherwise relational text strategies that add a global perspective to the text. A similar close reading analysis of the much less frequent clusters “alte welt” [old world] (32 occurrences) and “beide welten” [both worlds] (18 occurrences) further corroborated this thesis. Both of them appear exclusively in comparative contexts of global proportions.

---

32 A. v. Humboldt, *Reise*, vol. 2, 225. [A. v. Humboldt, *Personal Narrative*, vol. 4, 182.]

33 A. v. Humboldt, *Reise*, vol. 3, 172. [A. v. Humboldt, *Personal Narrative*, vol. 5, 141.]

34 A. v. Humboldt, *Reise*, vol. 4, 286. [A. v. Humboldt, *Personal Narrative*, vol. 6, 112.]



As a next step, I traced the word “erde” throughout the corpus which showed that it is actually used mostly as a non-spatial term. 86 of the 121 instances refer to the soil or ground of a specific place, for example in phrases such as “Scheiben aus gebrannter Erde” [“discs of baked earth”],<sup>35</sup> “schnee-weiße Erde” [“clay ... of a snowy whiteness”],<sup>36</sup> “eine Art Fort aus Erde und Holz” [“a kind of little fort, constructed of earth and timber”],<sup>37</sup> or “feuchte Erde” [“damp ground”],<sup>38</sup> rather than to the planet – meaning that in approximately 70 percent of the cases the term carried no spatial and especially no global meaning. That left only 35 instances of the term carrying a global connotation, for example in phrases such as “Umdrehung der Erde” [“rotation”, “the earth’s rotation”, “rotatory motion of the globe”, “rotation of the Earth”, “rotation of the globe”]<sup>39</sup> or “Krümmung der Erde” [“curve of the globe”, “curvature of the earth”, “rotundity of the Earth”].<sup>40</sup>

A search for “erdball\*” [globe] painted a slightly more complicated picture. The term does not occur very frequently either, it only appears 30 times in the corpus, but it consistently refers to the planet as a whole, addressing such things as “Geschichte des Erdballes” [“history of the globe”],<sup>41</sup> “Umwälzungen des Erdballes” [“revolutions of the Globe”, “revolutions of the globe”],<sup>42</sup> or “Verteilung der Arten auf dem Erdballe” [“distribution of various species on the globe”].<sup>43</sup> However, it is often set in a generic comparative context that stresses the regional diversity of the planet, as is best demonstrated in the following excerpt:

“Gewinnt man einen Ueberblick über die Geschichte unseres Geschlechtes, so sieht man diese Mittelpunkte antiker Kultur da und dort gleich Lichtpunkten

35 A. v. Humboldt, *Reise*, vol. 1, 120. [A. v. Humboldt, *Personal Narrative*, vol. 1, 279.]

36 A. v. Humboldt, *Reise*, vol. 2, 130. [A. v. Humboldt, *Personal Narrative*, vol. 3, 488.]

37 A. v. Humboldt, *Reise*, vol. 3, 202. [A. v. Humboldt, *Personal Narrative*, vol. 5, 206.]

38 A. v. Humboldt, *Reise*, vol. 4, 111. [A. v. Humboldt, *Personal Narrative*, vol. 5, 618.]

39 A. v. Humboldt, *Reise*, vol. 1, 23, 29, 32, and 33; vol. 3, 184. [A. v. Humboldt, *Personal Narrative*, vol. 1, 46, 58, 64, 65; vol. 5, 171.]

40 A. v. Humboldt, *Reise*, vol.1, 53; vol. 2, 38; vol. 4, 141. [A. v. Humboldt, *Personal Narrative*, vol. 1, 105; vol.3, 290; vol.5, 676.]

41 A. v. Humboldt, *Reise*, vol. 1, 68. [A. v. Humboldt, *Personal Narrative*, vol. 1, 133.]

42 A. v. Humboldt, *Reise*, vol. 1, 197; vol. 2, 57. [A. v. Humboldt, *Personal Narrative*, vol. 2, 273; vol. 3, 342.]

43 A. v. Humboldt, *Reise*, vol. 3, 189. [A. v. Humboldt, *Personal Narrative*, vol. 5, 181.]

über den Erdball verstreut, und gewahrt mit Ueberraschung, wie ungleich die Gesittung unter den Völkern ist, die fast unter demselben Himmelsstriche wohnen und über deren Wohnsitze scheinbar die Natur dieselben Segnungen verbreitet hat.”

[“In studying the history of our species, we see, at certain distances, these foci of ancient civilization dispersed over the Globe like luminous points; and we are struck by the inequality of improvement in nations inhabiting analogous climates, and whose native soil appears equally favoured by the most precious gifts of nature.”]<sup>44</sup>

On the one hand, the generic cultural comparison adopts a global perspective by generically referring to the “Erdball” [“Globe”] and to the “Geschichte unseres Geschlechtes” [“history of our species”]. On the other hand, it stresses the cultural differences between different regions of the world without naming these regions in particular. One of the few explicitly global terms of the corpus, so it appears, is often accompanied by a reference to the regional diversity of the planet.

The last term on my search word list further proved that comparative text strategies produce a global perspective even when the lexical features of the text carry strictly regional connotations. The term “erdstrich\*” [zone, part of the earth], which occurs 69 times in the corpus and lexically refers to specific regions of the planet, is seldom used in an exclusively regional context. In fact, according to my analysis the term is only used as a strictly regional term seven times. In all other cases, the contexts show a surprisingly consistent structure, as demonstrated in the following excerpts:

“Es ist bekannt, wie häufig die Leuchtwürmer in **Italien** und im ganzen **mit-täglichen Europa** sind; aber ihr malerischer Eindruck ist gar nicht zu vergleichen mit den zahllosen zerstreuten, sich hin und her bewegenden Lichtpunkten, welche im **heißen Erdstrich** der Schmuck der Nächte sind, wo einem ist, als ob das Schauspiel, welches das Himmelsgewölbe bietet, sich auf der Erde, auf der ungeheuren Ebene der Grasfluren wiederholte.”

---

44 A. v. Humboldt, *Reise*, vol. 4, 52. [A. v. Humboldt, *Personal Narrative*, vol. 5, 500.]

[“We know how common the glow-worm is in **Italy**, and in all the **south of Europe**; but the picturesque effect it produces cannot be compared to those innumerable, scattered, and moving lights, that embellish the nights of the **torrid zone**, and seem to repeat on the earth, along the vast extent of the savannahs, the spectacle of the starry vault of the sky.”]<sup>45</sup>

“Es ist auffallend, wie in den **heißesten** und in den **kältesten Erdstrichen** der gemeine Mann gleich sehr die Wärme liebt.”

[“It seems remarkable, that in the **hottest** as well as the **coldest climates**, people display the same predilection for heat.”]<sup>46</sup>

“Hier im **tropischen Erdstrich** wachen sie auf, wenn es wieder feuchter wird; dagegen in **Georgien** und in **Florida**, im **gemäßigten Erdstrich**, reißt die wieder zunehmende Wärme die Tiere aus der Erstarrung oder dem Zustande von Nerven- und Muskelschwäche, in dem der Atmungsprozeß unterbrochen oder doch sehr stark beschränkt wird.”

[“Here, in the **equinoctial zone**, it is the increase of humidity that recalls them to life; while in **Georgia** and **Florida**, in the **temperate zone**, it is the augmentation of heat, that rouses these animals from a state of nervous and muscular debility, during which the powers of respiration are suspended, or singularly diminished.”]<sup>47</sup>

“Die **nordischen Heiden**, die Steppen an **Wolga** und **Don** sind kaum ärmer an Pflanzen und Tierarten als unter dem herrlichsten Himmel der Welt, im **Erdstrich der Bananen und des Brotfruchtbaums**, 567000 qkm Savannen, die im Halbkreise von Nordost nach Südwest, von den Mündungen des Orinoko bis zum Caqueta und Putumayo sich fortziehen.”

[“The **heaths of the north**, the steppes of the **Wolga** and the **Don**, are scarcely poorer in species of plants and animals, than are twenty-eight thousand square leagues of savannahs, that extend in a semicircle from north-east to

---

45 A. v. Humboldt, *Reise*, vol. 1, 189. [A. v. Humboldt, *Personal Narrative*, vol. 2, 249.]

46 A. v. Humboldt, *Reise*, vol. 2, 233. [A. v. Humboldt, *Personal Narrative*, vol. 4, 196.]

47 A. v. Humboldt, *Reise*, vol. 3, 61. [A. v. Humboldt, *Personal Narrative*, vol. 4, 501.]

south-west, from the mouths of the Oroonoko to the banks of the Caqueta and the Putumayo, beneath the finest sky of the globe, and in the **climate of plantains and breadfruit trees.**”]<sup>48</sup>

The term “erdstrich” [zone, part of the earth] is usually used in a context of global comparison, again, adding a global perspective to the text. Methodologically, reading through these comparisons provided me with a number of new search words for a potential follow-up study, such as: zone\* [zone], breite\* [latitude], länge\* [longitude], \*west\* [west], \*ost\* [east], \*süd\* [south], \*nord\* [north], heiße\* [hot, torrid], kalte\* [cold], or gemäßigte\* [temperate]. The examples also pointed to the possibility of searching the corpus for specific names of countries or continents although such an analysis would require a lengthy process of preselecting possibly significant names – a process which in turn heavily depends on extensive knowledge of the text acquired through close reading methods. However, it is important to note that following the term “erdstrich” through the corpus revealed a rich practice of global comparing in the corpus which would not have been indicated by its lexical features.

In sum, the close reading analysis proved to be a crucial step to evaluate and complement the results of the distant reading process. First of all, it helped to determine the actual connotation of search terms as used in the corpus, for example when differentiating between the different meanings of the term “erde” [earth, soil, ground]. Beyond that it was necessary to trace comparative text strategies throughout the entire length of the corpus – text strategies that produce a sense of the global beyond the regional focus that the text displays on a lexical level.

## Using a reference corpus: degrees of globality

The observation that the corpus shows very different ways of engaging with the world on different levels raises some hermeneutical questions that call for further investigation. Are the lexical features of the corpus actually that significant when most engagement with the world happens on the level of text strategies? Can we come to a viable first interpretation of the corpus by

---

48 A. v. Humboldt, *Reise*, vol. 4, 264. [A. v. Humboldt, *Personal Narrative*, vol. 6, 70 f.]

artificially isolating these lexical features from their specific contexts as I did in the first step of this analysis? Comparing the results to a reference corpus confirmed the validity of such an approach, proving that even on a lexical level a corpus can indeed show varying degrees of engagement with the world.

I chose Humboldt's *Kosmos*<sup>49</sup> as a reference corpus, since the text explicitly focuses on a description of the world, on a "physische Weltbeschreibung," as indicated by the subtitle. That way I hoped to compare the travelogue which rather focuses on the voyage and specific places to a text that primarily focuses on the world in its entirety. The aim was to see whether the two corpora engaged with the world to a different extent according to their different areas of focus. As a first step, I used the keyword list to identify characteristic words in the *Reise*-corpus. Whereas the word list tool only counts the words in a corpus, the keyword list tool shows which words are unusually frequent or infrequent in comparison with the words in a reference corpus (in this case the *Kosmos*-corpus). In other words, the tool compares two corpora to each other, identifying keywords with regard to the frequency of the words in the two corpora.<sup>50</sup>

Using this tool for an analysis of the *Reise*-corpus showed that many of the terms that ranked highly in the initial word list reappeared in the keyword list, as can be seen in the following tables:

---

49 Humboldt, Alexander von, *Kosmos: Entwurf einer physischen Weltbeschreibung.*, 5 vols., Stuttgart/Augsburg: Cotta, 1845–1862. The text is loosely based on a number of lectures held in Berlin in 1827 and 1828. It addresses a wide variety of topics, such as cosmic nebulas, stars, volcanos, plant and animal life, or human history. Overall, it aims to give a survey of all the physical phenomena found in the world and in the universe.

50 For a more distinguished introduction to and reflection on key word analysis in historical corpora see Baron, Alistair/Rayson, Paul/Archer, Dawn, Word frequency and key word statistics in historical corpus linguistics, in: *International Journal of English Studies* 20 (2009), 41–67.

Table 4: Word list for the Reise-corpus & Table 5: Keyword list for the Reise-corpus

Word List Results 1				Keyword List Results 1				
Word Types: 35313		Word Tokens: 218084	Search Hits: 0	Keyword Types: 1679	Keyword Tokens: 94987		Search Hits: 0	
Rank	Freq	Word	Lemmas Word Form(s)	Rank	Freq	Keyness (LL4)	Effect (DICE)	Keyword
1	1107	orinoko		1	1107	+ 2065.38	0.0101	orinoko
2	950	rio		2	644	+ 1200.71	0.0059	walfer
3	746	m		3	950	+ 1175.29	0.0087	rio
4	661	großen		4	574	+ 939.08	0.0052	indianer
5	655	mehr		5	420	+ 782.81	0.0038	falt
6	644	walfer		6	390	+ 726.86	0.0036	km
7	574	indianer		7	350	+ 652.27	0.0032	küste
8	541	ganz		8	346	+ 644.81	0.0032	insel
9	524	luft		9	412	+ 610.9	0.0038	caracas
10	513	zeit		10	746	+ 590.28	0.0068	m
11	492	san		11	310	+ 577.69	0.0028	felbit
12	478	weit		12	337	+ 547.39	0.0031	negro
13	463	jahre		13	283	+ 527.35	0.0026	milffonen
14	447	boden		14	376	+ 508.43	0.0034	cumana
15	436	zwei		15	260	+ 484.48	0.0024	teil
16	420	falt		16	256	+ 477.02	0.0023	fieht
17	412	caracas		17	492	+ 466.81	0.0045	san
18	403	kleinen		18	254	+ 461.22	0.0023	gibt
19	390	km		19	245	+ 456.52	0.0022	reife

The initial word list already included many terms referring to spatiality, such as the names of specific rivers or places (“orinoko” and “caracas”), particles used in the names of rivers and places (“rio” and “san”), or metric markers (“m” and “km”). The keyword list confirmed that such a focus on local or regional space is specific to the travelogue, pointing to a number of terms that do not only occur frequently but can be considered as keywords in comparison to another corpus, in this case Humboldt’s *Kosmos*. High-ranking keywords included the aforementioned words but also others, such as “küste” [coast], “insel” [island], “cumana” [Cumaná], and “reise” [voyage], pointing to the travelogue’s specific focus on movement through regional space.

The second step was to swap the corpora, analyzing the *Kosmos*-corpus as the main corpus with the travelogue as the reference corpus. Both the word list and the keyword list rendered significant results:

Table 6: Word list for the Kosmos-corpus &amp; Table 7: Keyword list for the Kosmos-corpus

Word List Results 2				Keyword List Results 2				
Word Types: 51423		Word Tokens: 335387	Search Hits: 0	Keyword Types: 1432		Keyword Tokens: 113493	Search Hits: 0	
Rank	Freq	Word	Lemmas Word Form(s)	Rank	Freq	Keyness (LL4)	Effect (DICE)	Keyword
1	1020	schon		1	1020	+ 776.01	0.0061	schon
2	938	fuß		2	938	+ 629.81	0.0056	fuß
3	874	großen		3	621	+ 622.6	0.0037	fast
4	766	worden		4	628	+ 616.6	0.0037	kosmos
5	732	erde		5	544	+ 545.35	0.0032	vergl
6	692	a		6	606	+ 457.8	0.0036	planeten
7	635	mehr		7	449	+ 450.07	0.0027	erscheinungen
8	628	kosmos		8	443	+ 444.05	0.0026	theil
9	621	fast		9	377	+ 377.86	0.0022	ersten
10	606	planeten		10	370	+ 370.84	0.0022	erst
11	575	beobachtungen		11	366	+ 366.83	0.0022	theile
12	572	zeit		12	363	+ 363.83	0.0022	cometen
13	558	höhe		13	362	+ 362.82	0.0022	insel
14	557	ganz		14	345	+ 345.78	0.0021	zuerst
15	554	große		15	334	+ 334.75	0.002	herschel
16	544	vergl		16	376	+ 332.9	0.0022	cap
17	536	sterne		17	318	+ 318.71	0.0019	wahrscheinlich
18	535	jahre		18	329	+ 318	0.002	anm
19	530	sonne		19	310	+ 310.69	0.0018	sagt

In comparison to the travelogue and with regard to my interest in spatiality and globality, some of the high-ranking words can be considered especially significant. Both lists include a number of words referring to the universe and celestial bodies in particular. The word list refers to frequently occurring words, such as “kosmos” [cosmos, universe], “planeten” [planets], “sterne” [stars], “sonne” [sun], and “erde” [earth, soil, ground],<sup>51</sup> whereas the keyword list points towards similar terms that occur especially frequently in comparison to the travelogue, again including the words “kosmos” [cosmos, universe] and “planeten” [planets] but also adding “cometen” [comets] to the list. This lexical focus on cosmic space and celestial bodies seems almost complementary to the regional focus of the travelogue.

To test if this difference in focus affected the lexical use of the word “welt” [“world”] in the *Kosmos*-corpus, I searched for “welt\*” in the concordance and clusters tools. Again, the analysis revealed significant differences between the corpora. While the term in all its variations occurs 1306 times in the *Kosmos*, it mostly occurs in the form of those compound nouns that hardly ever occur in the travelogue:

51 Again, only a close reading analysis of the results could determine whether the term is actually used with reference to the planet or the soil. Nonetheless, even if the term had to be excluded from the results, the list would still be considerable.

Table 8: Results of the cluster analysis for “welt\*”, cluster size of one word (Kosmos-corpus) & Table 9: Results of the cluster analysis for “welt\*”, cluster size of two words (Kosmos-corpus)

Clusters Results 1				Clusters Results 2			
Total No. of Cluster Types 176			Total No. of Cluster Tokens 1306	Total No. of Cluster Types 623			Total No. of Cluster Tokens 1306
Rank	Freq	Range	Cluster	Rank	Freq	Range	Cluster
1	134	5	welt	1	37	5	physischen weltbeschreibung
2	133	5	weltkörper	2	32	5	der weltanschauung
3	79	5	weltbeschreibung	3	28	5	der weltkörper
4	67	5	weltanschauung	4	27	5	die welt
5	55	4	weltraum	5	26	3	den weltraum
6	37	4	weltraume	6	25	4	des weltraums
7	35	4	weltkörpern	7	23	4	im weltraume
8	27	5	weltall	8	22	3	der welt
9	27	4	weltraums	9	19	3	der vorwelt
10	26	2	weltherrschaft	10	19	5	im weltall
11	24	3	außenwelt	11	19	3	physische weltbeschreibung
12	21	3	weltansicht	12	18	2	physischen weltanschauung
13	20	4	tropenwelt	13	16	3	der außenwelt
14	19	3	vorwelt	14	14	3	der tropenwelt
15	17	3	inselwelt	15	14	2	des weltbaues
16	17	4	weltumsegung	16	13	3	neuen welt
17	16	3	weltganzen	17	12	4	des weltalls
18	16	4	weltordnung	18	10	2	dem weltraume
19	15	2	weltbaues	19	10	5	der gedankenwelt

A number of the high-ranking words refer to the cosmic or planetary dimension of the term “welt”, such as “weltkörper\*” [heavenly body], “weltraum\*” [space, outer space], or “weltall\*” [space, universe], while others refer to the term in a rather philosophical or political way, as in the cases of “weltanschauung\*” [world view], “weltherrschaft\*” [world domination], “weltansicht\*” [word view], or “weltordnung\*” [world order].<sup>52</sup> It is important to note that the terms which refer to the world in its totality also appear as most frequent ones when changing the cluster size to two words. The regional term “neuen welt” [new world], which is the highest ranking two word cluster in the travelogue, only ranks 16th with 13 occurrences in this corpus, seem-

52 The listed terms imply Ottmar Ette’s observation that Humboldt’s world concepts include both philosophical and planetary connotations. See O. Ette, *Humboldt und die Globalisierung*, 366 ff.; O. Ette, *Weltbewußtsein*, 90 ff. As mentioned above, compound nouns as these cannot be translated easily because they tend to carry very specific philosophical and epistemological connotations. The translations provided above can merely be considered suggestions that help the reader navigate through the analysis. To provide historically accurate translations, I would have had to track those terms through the different English translations of the time, which, with regard to the number of hits, would have exceeded the limits of this case study.



ing almost insignificant in comparison to the high-ranking compound noun clusters. These results show that the word “welt” is strongly associated with the universal and cosmic focus of the text, in contrast to the regional use of the term in the travelogue.

For a more comprehensive comparison of the two corpora, some further distant and close reading analysis of the above mentioned planetary terms, including the term “erde” [earth], would surely be necessary. However, even these preliminary results point to the hermeneutic productivity of comparing the distant reading results of two corpora. While previous research has strongly associated Humboldt’s claim to describe the world in its totality with specific ways of writing, such as descriptive, narrative, or comparative writing,<sup>53</sup> this analysis shows that even on a lexical level texts can engage with the world to different degrees. While the travelogue only displays such an engagement with the world on the level of comparative and relational text strategies, the lexical level displaying a focus on regional space, the *Kosmos*’s engagement with the world (and the universe) starts at a lexical level. It could be argued that such a corpus shows a higher degree of *Welthaltigkeit* since it approaches world concepts more explicitly. A genre such as travel writing appears to take a more implicit approach, achieving a global perspective through comparing regional entities to each other and thereby promoting a rather relational concept of the world.

Whether or not specific degrees of globality are characteristic for certain genres – travel writing is only one possible object of investigation – can only be decided by text mining much larger corpora than the ones in this case study. Nonetheless, in this case combining distant and close reading as well as comparing distant reading results concerning different corpora, has proven a productive way to ask about the different ways in which literary texts can engage with the world or refrain from doing so.

---

53 C. Peters, *Reisen und Vergleichen*; C. Peters, *Historical Narrative versus Comparative Description*; O. Ette, *Weltbewußtsein*, 158 ff.; Lubrich, Oliver, *Das Schwinden der Differenz: Postkoloniale Lektüren, Alexander von Humboldt – Bram Stoker – Ernst Jünger – Jean Genet*, Bielefeld: Aisthesis, 2004, 87 ff.; Kraft, Tobias, *Figuren des Wissens bei Alexander von Humboldt: Essai, Tableau und Atlas im amerikanischen Reisewerk*, Berlin/Boston: De Gruyter, 2014, 15 ff.

## Bibliography

- Anthony, Laurence*, A critical look at software tools in corpus linguistics, in: *Linguistic Research* 30 (2013), 141–161.
- Archer, Dawn*, Data Mining and Word Frequency Analysis, in: Gabriele Griffin/Matt Hayler (eds.), *Research Methods for Reading Digital Data in the Digital Humanities*, Edinburgh: Edinburgh University Press, 2016, 72–92.
- Baron, Alistair/Rayson, Paul/Archer, Dawn*, Word frequency and key word statistics in historical corpus linguistics, in: *International Journal of English Studies* 20 (2009), 41–67.
- Böhme, Hartmut*, Ästhetische Wissenschaft: Aporien der Forschung im Werk Alexander von Humboldts, in: Ottmar Ette (ed.), *Alexander von Humboldt: Aufbruch in die Moderne*, Berlin: Akademie-Verlag, 2001, 17–32.
- Daum, Andreas*, Alexander von Humboldt, die Natur als ‘Kosmos’ und die Suche nach Einheit: Zur Geschichte von Wissen und seiner Wirkung als Raumgeschichte, in: *Berichte zur Wissenschaftsgeschichte* 22 (2000), 246–250.
- Erhart, Walter*, Chamissos Weltreise und Humboldts Schatten, in: Julian Drews et al., *Forster – Humboldt – Chamisso: Weltreisende im Spannungsfeld der Kulturen*, Göttingen: V&R unipress, 2017, 13–34.
- Erlin, Matt*, Topic Modeling, Epistemology, and the English and German Novel, in: *Journal of Cultural Analytics* (2017).
- Ette, Ottmar*, Languages about Languages: Two Brothers and one Humboldtian Science, in: *HiN: Internationale Zeitschrift für Humboldt-Studien* XIX (2018), 47–61.
- Ette, Ottmar*, Humboldt und die Globalisierung: Das Mobile des Wissens, Frankfurt a. M./Leipzig: Insel, 2009.
- Ette, Ottmar*, Unterwegs zu einer Weltwissenschaft? Alexander von Humboldts Weltbegriffe und die transarealen Studien, in: *HiN: Internationale Zeitschrift für Humboldt-Studien* VII (2006), 34–54.
- Ette, Ottmar*, Weltbewußtsein: Alexander von Humboldt und das unvollendete Projekt einer anderen Moderne, Weilerswist: Velbrück, 2002.
- Ette, Ottmar*, The Scientist as Weltbürger: Alexander von Humboldt and the Beginning of Cosmopolitics, in: *HiN: Internationale Zeitschrift für Humboldt-Studien* II (2001).

- Ette, Ottmar*, Unterwegs zum Weltbewußtsein: Alexander von Humboldts Wissenschaftsverständnis und die Entstehung einer ethisch fundierten Weltanschauung, in: *HiN: Internationale Zeitschrift für Humboldt-Studien I* (2000).
- Fuchs, Anne*, Reiseliteratur, in: Dieter Lamping (ed.), *Handbuch der literarischen Gattungen*, Stuttgart: Alfred Kröner, 2009, 593–600.
- Görbert, Johannes*, Die Vertextung der Welt: Forschungsreisen als Literatur bei Georg Forster, Alexander von Humboldt und Adelbert von Chamisso, Berlin: De Gruyter, 2014.
- Heyl, Bettina*, Das Ganze der Natur und die Differenzierung des Wissens: Alexander von Humboldt als Schriftsteller, Berlin/Boston: De Gruyter, 2007.
- Humboldt, Alexander von*, *Personal Narrative of Travels to the Equinoctial Regions of the New Continent, during the Years 1799–1804*, 7 vols., translated by Helen Maria Williams, New York: Cambridge University Press, 2011.
- Humboldt, Alexander von*, *Reise in die Aequinoktial-Gegenden des neuen Kontinents*, vol. 1., translated by Hermann Hauff, Stuttgart: Cotta, 1859.
- Humboldt, Alexander von*, *Kosmos: Entwurf einer physischen Weltbeschreibung*, 5 vols., Stuttgart/Augsburg: Cotta, 1845–1862.
- Knobloch, Eberhard*, Alexander von Humboldts Weltbild, in: *HiN: Internationale Zeitschrift für Humboldt-Studien X* (2009), 34–46.
- Kraft, Tobias*, Erdwissen im Angesicht der Berge: Die Vulkanlandschaften der Jorullo-Ebene als Heuristik der Geologie, in: Ottmar Ette/Julian Drews (eds.), *Horizonte der Humboldt-Forschung: Natur, Kultur, Schreiben*, Hildesheim: Georg Olms, 2016, 97–124.
- Kraft, Tobias*, *Figuren des Wissens bei Alexander von Humboldt: Essai, Tableau und Atlas im amerikanischen Reisewerk*, Berlin/Boston: De Gruyter, 2014.
- Lubrich, Oliver*, Das Schwinden der Differenz: Postkoloniale Lektüren, Alexander von Humboldt – Bram Stoker – Ernst Jünger – Jean Genet, Bielefeld: Aisthesis, 2004.
- Moser, Christian/Simonis, Linda*, Einleitung: Das globale Imaginäre, in: Sebastian Moser/Linda Simonis (eds.), *Figuren des Globalen: Weltbezug und Welterzeugung*, Göttingen: V&R unipress, 2014, 11–22.
- Peters, Christine*, *Historical Narrative versus Comparative Description? Genre and Knowledge in Alexander von Humboldt's Personal Narrative*,

in: Martin Carrier/Carsten Reinhardt/Veronika Hofer (eds.), *Narratives and Comparisons: Adversaries or Supporters in Understanding Science?*, (manuscript in preparation).

Peters, Christine, *Reisen und Vergleichen: Praktiken des Vergleichens in Alexander von Humboldt's Reise in die Äquinoktial-Gegenden des Neuen Kontinents und Adam Johann von Krusensterns Reise um die Welt*, in: *IASL* 42 (2017), 441–455.

Stockhammer, Robert, *Welt oder Erde? Zwei Figuren des Globalen*, in: Christian Moser/Linda Simonis (eds.), *Figuren des Globalen. Weltbezug und Welterzeugung in Literatur, Kunst und Medien*, Göttingen: V&R unipress, 2014, 47–72.



# From Serial Sources to Modeled Data

## Changing Perspectives on Eighteenth-Century Court Records from French Pondicherry<sup>1</sup>

---

Anna Dönecke

A crucial point in historical research is the availability of sources, which is, especially for pre-modern times, not guaranteed. Hence, when I started my research on the French trading company in India during the eighteenth century, the issue of finding proper sources was a decisive point. However, as I wanted to focus on the jurisdiction in the French headquarter Pondicherry, it soon became clear that a dearth of sources would not pose a problem. The French administrators there produced an abundance of different documents, ranging from letters to the general directors in Paris and protocols of their meetings to a myriad of court records. Being enabled by the sources to delve deeply into the everyday proceedings of the jurisdictional field in Pondicherry soon proved to cut both ways. Although it may be presumptuous to speak of *big data* in comparison to the quantity of data natural scientists need to handle, the records provided a vast amount of historical data.<sup>2</sup> The task of processing all the information thus posed a major challenge, and at this point digital methods came into play in my humanistic research.

The best option to cope with the situation was to gather all available information in a relational database in a structured way which would also

---

1 I am very grateful to Antje Flüchter and Stephan Fasold for critical readings of the article.

2 For a discussion of *big data* and their handling in historical cultural studies cf. *Schmale, Wolfgang*, Big Data in den historischen Kulturwissenschaften, in: Wolfgang Schmale (ed.), Digital Humanities: Praktiken der Digitalisierung, der Dissemination und der Selbstreflexivität, Stuttgart: Franz Steiner, 2015, 137.

allow me to analyze them further. But to look at this in a merely output-oriented way and to conceive digital methods just as some kind of auxiliary tool for historians would be, as I want to show in this article, too short-sighted. Rather, their application itself or, in this case, the modeling of the database should be focused as an encounter of different ways of asking questions and handling source material which already can be productive. Like the digital humanist Willard McCarty puts it, the computational demand for complete explicitness and absolute consistency “effects a sea-change by forcing us to confront the radical difference between what we know and what we can specify computationally, leading to the epistemological question of *how we know what we know*.”<sup>3</sup>

Taking these observations as a starting point, this article focuses rather on the *making* of the database for my project than the database itself. It takes a humanistic point of view and conceives the modeling of data as “a constructive and creative process”<sup>4</sup> starting with unwieldy information in a bulk of sources and ending with a formal model representing this information in a structured way. To do so, I focus on two questions while tracing the process of modeling the database for my project. On a general level, I ask what it means for a trained historian to apply digital methods in her research and point out some specific difficulties arising from historical sources. Furthermore, I examine the way in which the necessity of developing an abstract grid for the database which grasps all relevant information of the source retroactively affects how one conceives the sources.

I will begin with a brief depiction of my research project, followed by a description of the database, its specific requirements, and the reason for its choice. Then, I will track the modeling of my data in three sections: In the initial step, I reconsider the sources and derive a conceptual model from them, which is then transferred into a structure for the database, before I concern myself with the issue of importing data into the database.

---

3 McCarty, Willard, *Humanities Computing*, Basingstoke: Palgrave Macmillan, 2014, 25.

4 Flanders, Julia/Jannidis, Fotis, *Data Modeling*, in: John Unsworth/Raymond George Siemens/Susan Schreibman (eds.), *A New Companion to Digital Humanities*, Chichester, West Sussex, UK/Malden, MA, USA: Blackwell, 2016, 234.

## Point of departure: intercultural jurisdiction in eighteenth-century Pondicherry

My research project<sup>5</sup> on the intercultural jurisdiction in Pondicherry draws on the observation that, upon its arrival at the Coromandel Coast, the *Compagnie des Indes Orientales* was by no means in the position to dictate its terms. Rather, it found itself in a multipolar constellation of power, even after acquiring Pondicherry in 1674 from Sher Khan Lodi, the region's ruler. Not only was it competing with other European powers present in South India but it was also dependent on the different regional rulers' benevolence and relying on the local society's cooperation to establish a flourishing trading post. In this situation, the jurisdiction in Pondicherry was a central field of interaction where members of the different groups encountered one another. Also, it was an arena in which conflicts between individuals were decided as well as power relations between the French and local groups negotiated.<sup>6</sup>

In my project, I examine this process of negotiation within and on the jurisdictional field with a particular focus on the agency of local groups in

---

5 I am conducting this research as PhD project within the framework of the Collaborative Research Center SFB 1288 "Practices of Comparing. Changing and Ordering the World", Bielefeld University, Germany, funded by the German Research Foundation (DFG), sub-project B01 "Order in diversity: Practices of comparing in intercultural jurisdiction (17th–19th century)" lead by Antje Flüchter and Christina Brauner.

6 This perspective should by no means imply harmonic conditions or neglect asymmetric constellations in situations of cultural contact. As Christina Brauner and Antje Flüchter emphasize in their article on state-building in a transcultural context: "For rule and governance to be established in a functional manner in a contact zone, the different cultural (in this case mostly governmental) routines have to be made compatible. [...] Negotiation (in a broad sense) [...] can take place in the context of different power constellations, from more or less balanced relations to situations of striking power asymmetries." Brauner, Christina/Flüchter, Antje, Introduction: The Dimensions of Transcultural Statehood, in: Christina Brauner/Antje Flüchter (eds.), *The Dimensions of Transcultural Statehood*, Leipzig: Leipziger Universitätsverlag, 2015, 23. Cf. also Flüchter, Antje, Structures on the Move: Appropriating Technologies of Governance in a Transcultural Encounter, in: Antje Flüchter/Susan Richter (eds.), *Structures on the Move: Technologies of Governance in Transcultural Encounter*, Berlin/London: Springer, 2012, 1–27; Parasher, Gauri, Between Sari and Skirt: Legal Transculturality in Eighteenth-Century Pondicherry, in: Christina Brauner/Antje Flüchter (eds.), *The Dimensions of Transcultural Statehood*, Leipzig: Leipziger Universitätsverlag, 2015, 56–77.



the light of increasingly prominent attempts to expand French control in Pondicherry during the eighteenth century. Therefore, I call into question contemporary claims of the French to legal sovereignty in their Indian trading post and ask for historical shifts in this context during the century which is commonly identified as the beginning of European or British colonialism in India. I assume that historical actors relied on practices of comparing to maneuver themselves through the pluralistic situation and to handle its legal and cultural diversity, and that these practices thereby played an important role in establishing a legal order in Pondicherry.<sup>7</sup> In particular, I am interested in how the different legal authorities – French and ‘indigenous’ – interacted with each other. Moreover, I ask how the different actors – litigants and judges – handled the co-existence of local as well as French legal orders and institutions, and pay particular attention to their practices of comparing.<sup>8</sup>

To do so, I draw upon a wide range of sources. To a large proportion the material consists of different types of documents from the *Conseil Supérieur de Pondichéry*, the highest French court in India. Those are mainly court records, correspondence, and regulations regarding the jurisdiction in Pondicherry. To broaden the perspective, I also use ego documents, for example letters from the administrators, memoirs from *Compagnie* employees or travelogues. Furthermore, the records of the *Tribunal de la Chaudrie* are of special importance. The *Chaudrie* functioned as a court for all local groups in civil matters and became progressively controlled by the French authorities in the course of the eighteenth century.<sup>9</sup>

---

7 For more information on practices of comparing, cf. *Epple, Angelika/Erhart, Walter (eds.), Die Welt beobachten: Praktiken des Vergleichens*, Frankfurt a. M.: Campus, 2015.

8 For a more elaborate account and preliminary results of my project, see my article *Dönecke, Anna*, ‘Le chapeau ou la toque’: Rechtliche Vielfalt und soziale Diversität in Pondichéry im 18. Jahrhundert, in: *Christina Brauner/Antje Flüchter (eds.), Recht, Ordnung, Diversität*, Bielefeld: transcript, forthcoming.

9 For an introduction to the legal landscape of Pondicherry in the eighteenth century see *Houllemare, Marie*, La justice française à Pondichéry au XVIIIe siècle, une justice en ‘zone de contact’, in: *Éric Wenzel/Éric de Mari (eds.), Adapter le droit et rendre la justice aux colonies: Thémis outre-mer, XVIe-XIXe siècle*, Dijon: Éditions Universitaires de Dijon, 2015, 147–157.

## Choosing a digital method: modeling a relational database

Although it is clear by now that my engagement with Digital Humanities served a specific research purpose, it is nevertheless necessary to stress this point when talking about the selection of a proper method or tool. As Julia Flanders and Fotis Jannidis put it, researchers in comparison to archivists or librarians “typically concentrate on producing data that will be more specifically directed towards their own research needs”<sup>10</sup>. Consequently, the research need is the guiding principle for choosing a proper method and tool.

Since I am especially interested in cases involving locals as well as Frenchmen, the *Chaudrie* records take up a central role in my research and thus served as the starting point. A part of them was edited and published by Jean-Claude Bonnan in two volumes, roughly comprising 250 cases in total.<sup>11</sup> These records are available as PDF scans and could therefore be handled by means of digital methods. First of all, the scans were transformed into machine readable text via OCR.<sup>12</sup> The records follow the same structure; all provide certain information at a specific position in the text with standardized phrasing: A header with three lines was added in the course of the publication, indicating the date of the trial, the name of the litigants, and their social affiliation, respectively. Hence, it was possible to computationally extract specific information by using regular expressions and gather them in a comma-separated values (CSV) file. In the CSV file, each line represented a case and consisted of several comma-separated fields that contained different information on it. As a result, the data was organized in a structured but still linear way, following the order of its appearance in the sources, and the options to evaluate it were very limited. At this point, a database seemed the most suitable tool to work with the collected data and to enable me to even

---

10 J. Flanders/F. Jannidis, *Data Modeling*, 233.

11 Bonnan, *Jean-Claude* (ed.), *Jugements du Tribunal de la Chaudrie de Pondichéry. 1766–1816*, 2 vols., Pondicherry: Institut français de Pondichéry, 1999. For more information on this, cf. *Menski*, *Werner*, Jean-Claude Bonnan, *Jugements du tribunal de la chaudrie de Pondichéry 1766–1817*, in: *Indo-Iranian Journal* 46 (2003), 369–371. Yet, Bonnan's edition only comprises a small part of the records. A lot more survived, partly due to copies that were made in the nineteenth century. They are nowadays hold in the *Archives nationales d'outre-mer* in Aix-en-Provence.

12 For this, the OCR pipeline of the INF project was used. For more information, cf. the contribution by Patrick Jentsch and Stephan Porada in this volume.

add further findings from readings of the *Chaudrie* or the *Conseil* records. Let me illustrate this in more detail by drawing a comparison.

The comma-separated data can best be envisioned as a spreadsheet with each row containing information on a specific trial and the columns stating the date, the plaintiff's and the defendant's names, and potentially additional data like the judges' names, respectively. Obviously, this spreadsheet contained redundancies – judges were involved in more than one trial, of course, thus their names would appear multiple times on the spreadsheet. More importantly, there would only be a few options to manipulate the data. The lines could be arranged in a chronological order, sorted alphabetically by the names of the plaintiffs or by the number of occurrences of a specific name. But since the object of study in my research is foremost the *interaction* of different actors and authorities in Pondicherry's jurisdiction, more advanced options to record and interrogate the data were needed.

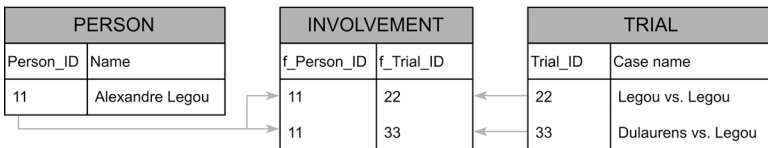
Quamen and Bath even choose this aspect as the defining point of databases: "A database, according to our definition, is a rigorously organized set of data whose informational patterns help us maximize the number of possible questions we can ask of it." They point out that "if you are embarking upon a project in which you will be actively engaging with your data, pushing its limits, and asking challenging questions of it – finding patterns, seeing changing dynamics over time, locating anomalies, looking for missing information – then you will need a database."<sup>13</sup> In order to stick to the notion of a spreadsheet, one can think of relational databases – the most common type – as multiple spreadsheets that are connected to one another. Apart from entity tables for, say, trials and persons, they also contain so-called junction tables that manage relationships between entities. In this example, they would match persons and trials together and, thus, capture information on the involvement of persons in trials. The mechanism by which this is done is identifying each set of data by a unique identifier (ID), called primary key, and using this ID, then called foreign key, to refer to it in other tables.

---

13 Quamen, Harvey/Bath, John, Databases, in: Constance Crompton/Richard J. Lane/Ray Siemens (eds.), *Doing Digital Humanities: Practice, Training, Research*, London/New York: Routledge, 2016, 181.

For example, as shown below, Alexandre Legou would be identified as 11 in the table ‘person’ while the inheritance dispute he was involved in would be put into the table ‘trial’ with the ID 22. Those sets of data would then be connected in a junction table called ‘involvement’ by matching 11 and 22 as foreign keys. If Alexandre Legou were involved in an additional dispute tagged with the key 33, one would create a new entry in the ‘involvement’ table using 11 again to refer to the very same data set in the ‘person’ table and connect it with 33. By following these primary-key-to-foreign-key links, a relational database manages relationships without creating redundancies.

Fig. 1: Example for tables connected via IDs



This distribution of data between multiple spreadsheets – the “atomization of data”<sup>14</sup> – and their interconnectedness increase the number of potential questions that can be posed to the data. Especially its ability to capture relationships between different entities made a relational database the most fitting tool for my research. It provided the possibility to organize my data in a way that allows me to pose questions that are directed at the entanglement of different actors and institutions on the legal field and to ask for diachronic changes. As a tool to build such a database, the database management system *FileMaker* was chosen because it not only provides the possibility to design a customized interface for possibly entering further data later on but also a tool to import CSV data.

## From text to modeled data

As mentioned before, I want to focus on the modeling of data as a constructive process. That means neither the sources nor the data can be regarded as an adequate description of real historical entities but rather as an interpre-

<sup>14</sup> H. Quamen/J. Bath, *Databases*, 184.

tation (of an interpretation) of them. The sources already represent a subjective perspective on historical events and by no means describe them as they ‘really’ took place. Moreover, the historian actively creates the set of sources he or she works with by choosing the most fitting for his research interest and leaving others out. This ‘alienation’ is even taken one step further when the scholar singles out specific aspects of the sources’ abundance of possible information in the course of his or her research.

Taking this into account, modeling data can be understood as a “process of abstraction”<sup>15</sup> that starts with sources referencing historical entities and ends with a rather abstract or formal depiction that brings their information into line with computational requirements. In short, it is all about reducing and enclosing the ‘unwieldiness’ of data historians come across in their sources. During this process, it is crucial to keep in mind “that the formalized model determines which aspects of the subject will be computable and in what form”.<sup>16</sup>

### **a. Reconsidering the sources: conceptual data modeling**

Since the sources serve as the very basis not only for the automatic extraction of data but consequently also for the database, they are the starting point. So, the initial question to ask is: What information do they actually contain and which is relevant to my research purpose? The judicial records from Pondicherry show a wide range of different conflicts between individuals and contain a lot of detailed information about the trials. Yet, they describe a social situation with relatively fixed rules or roles. In each trial, there are a plaintiff and a defendant arguing about an issue that is finally decided upon by judges. In addition, court records even in the eighteenth century followed more or less specific rules or conventions. Those not only determined how they were written but also which information was included in the text. Hence, the records are structured similarly and contain the same categories of information. For example, very detailed information is given in most records about the written procedure, like the dates when requests were submitted by the litigants or the number with which the protocol was registered. Also, as different laws for different social groups existed, the social

---

15 J. Flanders/F. Jannidis, *Data Modeling*, 230.

16 *Ibid.*, 229.

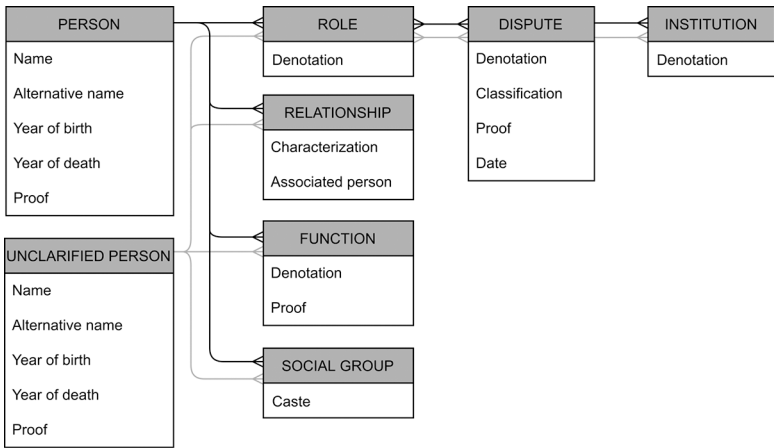
affiliation or, to be precise, the ascribed social affiliation of the litigants was usually mentioned at some point in the record. As the disputes were often about family matters like inheritance issues, one often finds information about family relationships between the persons involved. Moreover, it was quite common for the *Chaudrie* to rely on local headmen, the so-called *chefs de caste*, who functioned as arbitrators in smaller conflicts or as a court of appeal if the litigants were not satisfied with their arbitral decision. Consequently, its records often show traces of the interaction with them. Similarly, people could appeal to the *Conseil Supérieur* against the *Chaudrie*'s decisions. Finally, the *Chaudrie*'s records as well as those of the *Conseil* attained legal validity through the judges' signatures and thus, all end with the judges' names. However, not all of these details are immediately relevant to my research. As I am interested in Pondicherry's legal and social order and focus on the historical actors and their interaction, the following aspects could be singled out: Firstly, the personal information about the actors, i. e., their names, the denotation of their social affiliation, their role in the trial or their official function, their relationships to one another. Secondly, information about the trial itself is of interest, i. e., its date and the legal fora or institutions involved.

The next step is to convert these observations into a conceptual model. Here, a "purpose-oriented depiction is created that extracts a manageable amount of entities, attributes and relationships out of the plenitude of real world information".<sup>17</sup> The aim is to grasp the structure of the source's information and through this first abstraction prepare a basis for the actual database model. This is most commonly realized by means of an entity relationship diagram that can be understood as a visualized answer to the questions: What entities can be identified, what attributes do they have and how are they related to each other?

---

17 Jannidis, Fotis, Grundlagen der Datenmodellierung, in: Fotis Jannidis/Hubertus Kohler/Malte Rehbein (eds.), Digital Humanities: Eine Einführung, Stuttgart: Metzler, 2017, 103. Original quote in German: "In der konzeptuellen Modellierung wird eine zweckgebundene Abbildung aus der Fülle realweltlicher Informationen auf eine überschaubare Menge von Entitäten, Attributen und Relationen erstellt."

Fig. 2: Conceptual model (together with Anna Maria Neubert from the INF team)



Each box in this diagram represents an entity with its attributes while the lines connecting the boxes are used to depict the types of relationships these entities share. The trident side of the line signifies ‘many’ while the straight end represents ‘one’.

In the first place, there were seven entities to be distinguished in the court records: persons, the role they play in the trial, the dispute or trial itself, the institution or court, relationships the people had with one another, the office or function the persons held, and their social affiliation. The entity ‘unclarified person’ points towards specific aspects that one encounters when working with historical sources. Names were often spelled very differently, even if the same person was meant. Also, Indian converts were sometimes referred to by their Christian name, sometimes by their Indian name. Moreover, sons often bear the name of their fathers. As a result, it is sometimes – even if drawing on multiple additional sources – not possible to ultimately distinguish them for sure. For example, it is not possible to know whether Lazar and Tãnappa Mudali<sup>18</sup> or Jacques Dulaurens who was

18 This example is taken from the description of the inheritance dispute that evolved after the death of Kanakarāya Mudali in 1747. See Pillai, *Ananda Ranga*, The Private Diary of Ananda Ranga Pillai: Dubash to Joseph François Dupleix [...] A Record of Matters Political, Historical, Social, and Personal from 1736 to 1761, ed. and transl. by]. Frederick Price, Madras: Government Press, 1904, vol. 1, 310–375 and *Anonymus*, Arrêt du 20.03.1747, in:

appointed secretary at the *Conseil* in 1718 and Jacques Dulaurens appearing as a judge at the *Conseil* in a record from 1739<sup>19</sup> were two different persons or one and the same man.

Secondly, relevant attributes – again with regard to my research – were added to each entity. While attributes like ‘name’ or ‘caste’<sup>20</sup> are derived more or less directly from the sources, other attributes were added due to the database’s purpose. The attribute ‘proof’ is needed to meet the requirement of traceability. Aside from the necessity to keep track of the sources that have already been processed, one also needs to be able to give proof where the information was found. The attribute ‘classification’ is located at an analytic level, meaning that it is me as a scholar undertaking it. It already anticipates the possible need to sort out different types of disputes, for example inheritance or adoption issues, when analyzing the data in the end.

Thirdly, the diagram also represents the relationships the different entities share. A person usually has multiple relationships, being for example a father and a brother at the same time. Also, a person can have different functions: The judges of the *Chaudrie* were mostly members of the *Conseil Supérieur* at the same time. Although rather unusual, a person could be assigned to more than one social group. For instance, an Indian convert could be described as a Christian while still being seen as a member of the pariahs. A person could also be involved in more than one dispute and play different roles in each. For example, he or she could be the plaintiff in an inheritance dispute while being sued for not paying his or her debts in another trial. Lastly, there could be more than one institution involved in resolving a dispute. People could for example try to appeal to the *Conseil Supérieur* when

---

Gnagou Diagou (ed.), *Arrêts du Conseil Supérieur de Pondichéry*, Pondicherry: Société de l'histoire de l'Inde française, 1935, vol. 1, 178–81.

19 *Anonymus*, Procès-verbal du 22.06.1718, in: Edmond Gaudart (ed.), *Procès-verbaux des délibérations du Conseil Supérieur de Pondichéry*, Pondicherry: Société de l'histoire de l'Inde française, 1913, vol. 1, 191–192 and *Anonymus*, Arrêt du 31.03.1739, in: Gnagou Diagou (ed.), *Arrêts du Conseil Supérieur de Pondichéry*, Pondicherry: Société de l'histoire de l'Inde française, 1935, vol. 1, 58–59.

20 It should be noted here that the notion of a pan-Indian caste system as the defining feature of Indian society has been debated in the past decades. One widely shared result is that it was indeed implemented in the nineteenth century under the British Raj. Cf. for example *Dirks, Nicholas B.*, *Castes of Mind: Colonialism and the Making of Modern India*, Princeton, New Jersey: 2001. Nevertheless, the French frequently used the term caste in their records and classified the litigants by their (alleged) caste affiliation.

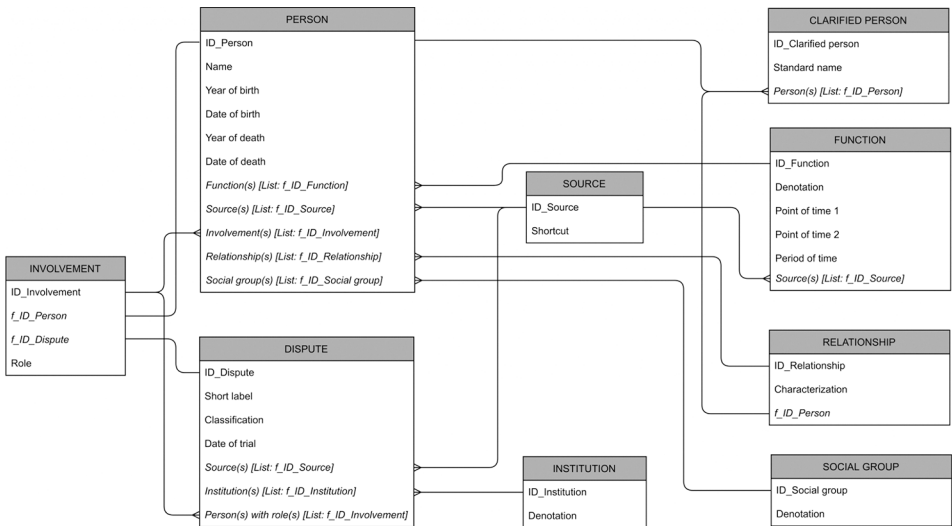


they were not satisfied with the *Chaudrie's* decision or, more common, the *Chaudrie* referred a case to the caste chiefs.

### b. Designing the database: logical data modeling

The next step on the path to a database is to transfer the entity relationship diagram, which represents the structure of the source’s relevant data, into a final database structure: “While the *conceptual* model has its origins in structures of meaning, the emphasis of the *logical* model is on providing a structure for the data that allows the user to use a set of algorithms to answer questions of interest in relation to the data.”<sup>21</sup> This, again, can be best visualized as a set of interconnected tables.

Fig. 3: Logical model (in cooperation with Silke Schwandt and Patrick Jentsch from the INF team)



At the center of this structure, there are the two master tables: ‘personal data’ and ‘dispute’. These are the main entities identified before, all other entities are linked to either one or both of them in the entity relationship diagram. Persons and disputes are connected via IDs in a junction table called ‘involve-

21 J. Flanders/F. Jannidis, Data Modeling, 231. Emphasis added.

ment' that captures which persons were involved in which disputes playing which role. Behind the two master tables reside several other tables containing additional information regarding 'personal data' and 'disputes'. They are connected to the two master tables through IDs.

As we have now reached the step which is about modeling the data in a way that finally meets computational requirements, specific challenges arise. Above all, a way to handle the uncertainty of data has to be found. The first problem already announced itself in the last step: the impossibility of always clearly identifying persons. In this case, it becomes necessary to allow potential redundancies in the 'personal data' table. Here, all persons appearing in the sources are gathered. For example, we would create a set of data for both, Lazar *and* Tânappa Mudali. If, by reading the source more carefully or consulting additional sources, I found out they were basically the same person, I would connect both sets of data to one another in the junction table 'clarified person'. Secondly, there is the possibility of incomplete or partial data. Sometimes, when looking for the date of birth or death of a person, all one can figure out is the year in which she or he was born or died. Finding the exact date is rather the exception. To cover both options, different fields were created in the 'personal data' table: year and date of birth or death. Thirdly, it is often unclear when exactly a person acquired or lost a function. In many cases, it is only possible to look for his (there were no female office holders) first and/or last appearance in the sources in a specific function. Also in this case, the fields were created accordingly: 'period of time', 'point of time 1', and 'point of time 2'.

Another challenge in this step was the removal of the 'logical' problems from the entity relationship diagram. As redundancies are unwanted in a database, the repetition of the 'proof' field throughout the entities or tables had to be avoided. Without a proper solution, the same source would have appeared numerous times, since a single record usually contains data on different entities and fields. By creating an extra table for sources and tagging each of them with an ID, it became possible to refer to sources in the 'proof' fields across the different tables by using a foreign key. Lastly, there was a many-to-many-relationship in the diagram between 'role' and 'dispute' because a person could be involved in different roles in more than one dispute. This had to be resolved into two one-to-many relationships stored in a junction table. Since a person could only obtain one role in a dispute, the former entity 'role' was transferred into an attribute and field in the junc-

tion table ‘involvement’, now denoting the role a person played in a particular trial.

The thus structured data allows for a series of challenging questions. One set is directed at proportions which can be found in the trials of the *Chaudrie*: What can be said about the distribution of litigants in terms of their social background? To what extent were intra-caste and inter-caste conflicts brought before the *Chaudrie*? What kinds of conflicts were mostly brought before the *Chaudrie*? In which amount of conflicts were the so-called caste chiefs involved? The second set of questions asks for diachronic developments and is designed to verify preconceived hypotheses. As mentioned before, one can observe from the sources that the French set out to expand and tighten their control over the jurisdiction in the second half of the eighteenth century. In this context, they also drew a sharper line between the *Conseil Supérieur*, which had been a court for the European population, and the *Chaudrie* by restricting access to the former for the local groups. The modeled data now allows reassessing the practical effects of this undertaking because it not only captures the interaction of different institutions in one and the same trial but also the dates of the trials. Hence, it is possible to ask if the frequency of appeals before the *Conseil Supérieur* against judgments from the *Chaudrie* changed over time. Furthermore, one is enabled to ask whether there was an increase in the involvement of lawyers in the disputes which could indicate a process of professionalization of legal counseling in the course of the century.

### c. Question of practicability: data import

The decisive step after modeling the database is filling it with actual information from the sources. The idea was to use the automatically extracted CSV data and to import them into *FileMaker*. But before the actual import could start, a closer look at the CSV file was necessary. It soon becomes obvious that it contains a certain amount of errors. While the dates of the trials were identified correctly in all instances, some errors occurred regarding the litigants’ names. They are normally separated by “c/” which stands for “contre” in the records. This made it possible to split them into parties: “participant\_a” and “participant\_b”. As there are nonetheless a few deviations from the standardized phrasing that served as the basis for the regular expression, this could not always be handled correctly by the machine as one can see in the chart below.

Fig. 4: Screenshot of the CSV file showing exemplary data sets (opened in LibreOffice Calc)

	A	B	C	D	E	
1	participant_a	participant_b	iso_date	date	participants	case_text
2	Kader Sahib	Mirabaté	1766-11-11	Ma 11 novembre 1766	Kader Sahib c/ Mirabaté	Choulia. En droit islamique
3	François et	No correct split.	1766-12-09	Ma 9 décembre 1766	François et	Carachi Chrétiens, Topas.
4	Chinoudou	Dobascayen	1766-12-09	Ma 9 décembre 1766	Chinoudou c/ Dobascayen	Malabar pa'en et Malabar
5	Andipa (ou Andiapa)	Chavrimoutou	1766-12-19	Ve 19 décembre 1766	Andipa (ou Andiapa) c/ Chavrimoutou	Le droit de revendiquer un
6	Tanapen	Levé	1767-01-23	Je 23 janvier 1767	Tanapen c/ Levé	Maure et Pally. Un père pe
7	Mahamadoussaib	Chinatamby	1768-04-11	Lu 11 avril 1768	Mahamadoussaib c/ Chinatamby	Musulman et Chetty. Une
8	Chinapen	Savamanden	1769-03-02	Je 2 mars 1769	Chinapen c/ Savamanden	Le tribunal peut contraindre
9	Cavqué Sinnapen	Samanden	1769-03-04	Sa 4 mars 1769	Cavqué Sinnapen c/ Samanden	La vente sur saisie d un ir
10	Tailamé	Calati	1769-04-04	Ma 4 avril 1769	Tailamé c/ Calati	Les frais d entretien d un t
11	Rangapa Soncoupam	Chavraia	1769-04-28	Ve 28 avril 1769	Rangapa Soncoupam c/ Chavraia	Le tribunal autorise une é
12	Ariatal vve Arlapen	No correct split.	1773-03-23	Ma 23 mars 1773	Ariatal vve Arlapen	Pally, chrétien. La conditio

For example, there are cases in which the litigants are a group of people, like the one represented in row 3 where the original text says “François and Jacques Tarabellion c/la soeur de Marie Perera veuve Luc Carachi”.<sup>22</sup> Here, another problem is already indicated: Female litigants are often not explicitly named but only referred to by means of their relation to male family members. Sometimes a short remark about their family status, e. g., their widowhood, is added, as can be seen in the case of “Ariatal v[eu]v[e] Arlapen” in the last row. This example reveals another particular problem: The lack of information on the second party of the dispute in the header or the absence of the “c/” which also leads to the output “No correct split”. Apart from that, there are more random deviations like in the record represented in line 5 where there is expressed insecurity about the spelling of the litigant’s name: “Andipa (ou Andiapa)”. However, the overall error rate is rather low, and the correct information is readily accessible in the records’ header. The problem of corrupt data could thus be solved without expending too much effort by manually post processing the file.

At least in theory and as initially planned, these data could now be easily imported into the *FileMaker* database. However, during this practice, one is confronted with the complicated issue of matching the CSV data to the database structure. It has been clear from the outset that some tables like ‘clari-

22 *Anonymus*, Tarabellion c/veuve Luc Carachi, 09.12.1766, in: Jean-Claude Bonnan (ed.), *Jugements du Tribunal de la Chaudrie de Pondichéry. 1766–1816*, 2 vols., Pondicherry: Institut français de Pondichéry, 1999, vol. 1, 3.

fied person' and 'relationship', and fields such as 'date/year of birth/death' in the database can only be filled manually while carefully reading the sources. This, unfortunately, also concerns the 'social group' of the litigants. Other than expected, the respective data could not be extracted from the records automatically because of their high degree of variation in this respect.<sup>23</sup> But most notably, there were several fields whose required information could be easily derived or generated from the CSV file but was not explicitly stored in it and thus could not be imported without further editing. Revisiting the distinction between information on entities and relationships, one can distinguish two different though closely related difficulties and approaches to solve them.

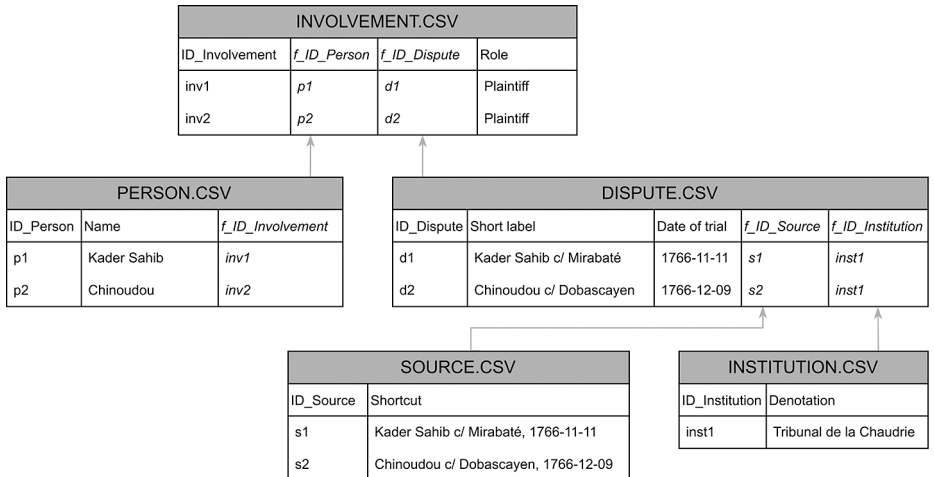
Firstly, entity data was missing in the CSV. With respect to the 'institution' table and its 'denotation' field, this could be solved easily because all evaluated records report on trials that took place before the *Tribunal de la Chaudrie*. Hence, a column entitled 'institution' was added to the CSV file, and its lines were all filled with "Tribunal de la Chaudrie". The values for the database's table 'source' could also be generated within the CSV file. The 'shortcut' field needed to distinctly reference the sources. As the records are sorted chronologically in the edition and there is not more than one trial with the same litigants and the same date, one could simply merge the values 'participants' and 'iso\_date' in a new column. For example, the dispute between Kader Sahib and Mirabaté which took place on November 11th 1766 would be referenced as "Kader Sahib c/ Mirabaté, 1766-11-11". A bit more challenging but still manageable was the creation of values for the 'role' field in the junction table 'involvement'. This field refers to the particular role of persons involved in a specific dispute. As the plaintiff is always mentioned before the defendant, one can derive this information from the 'participants' as well as from the 'participant\_a' and 'participant\_b' values. Two new columns had to be added to represent these pieces of information in a form that is interpretable by *File-Maker*. In the first column entitled 'names', the values from 'participants\_a' and 'participants\_b' were compiled underneath each other. The second column named 'role' was accordingly filled with "Plaintiff" and "Defendant".

---

23 There is a not too small number of records where the social affiliation is not mentioned in the header. Also, the parties could both belong to the same social group or to different groups. Sometimes, even a single person can be ascribed to different groups as mentioned before.

Secondly, as described above, the defining advantage of a database is its ability to manage relationships between different entities, and the CSV file does not meet the corresponding requirements. To illustrate the problem in more detail, the notion of a database as a set of interconnected entity and junction tables can once again be useful. The CSV file, on the contrary, can be viewed as a mere entity table. While one knows that the values in one line all refer to the same trial, there is no explicit information at hand about this relationship. Consequently, when it comes to the *FileMaker* import, no values are provided for the junction tables which connect the different entities via IDs. Thus, respective IDs need to be added to the CSV file in order to represent the relationships between the different entities as defined in the logical model. Therefore, new CSV files corresponding to the tables defined in the logical model were created on the basis of the original CSV file. The table's field here became columns that were successively filled with the extracted (and corrected) data. Most importantly, an identifier for each set of data was generated (ID\_Involvement, ID\_Person, ID\_Dispute, ID\_Source, ID\_Institution). These IDs were then copied to the other tables' corresponding columns and subsequently functioned as foreign IDs as illustrated below.

Fig. 5: Depiction of the different CSV files referencing one another by IDs



After this had been accomplished, there was explicit information on the different relationships between the entities in the form of IDs which could be interpreted by *FileMaker*. The CSV data finally matched the database structure which rendered it ready for its import.

## Conclusion

The emphasis of this article was put on the way towards a historical database, seizing on a central observation from Willard McCarty that “models of whatever kind are far less important to the digital humanities than *modelling*. Modelling is crucial.”<sup>24</sup> The main point of interest was the examination of the application of digital methods in historical research from a humanistic point of view.

It has become clear that through successively ‘atomizing’ the information from the court records into entities with attributes and relationships between one another, the sources’ information could finally be transferred into a data structure that could be surveyed computationally. Also, some practical difficulties were laid bare in the process. Firstly, the uncertainty or partiality of data historians are used to work with emerged as a problem when facing computational demands. This could be solved by incorporating those factors into the data structure. Secondly, the approach to automatically import the data from the sources into the database turned out to bear unanticipated. Besides the need to manually correct corrupt data, the diverging structures of the CSV data and the database posed a challenge: The requirements for the data to be imported rise with the complexity of the evaluation potential of the database. On the one hand, the automatic extraction of data lightened the workload to a great extent; on the other hand, its results could only be used further if one was able and willing to invest further time in its post processing. At this point, one might see the limits of digitally assessing texts. But, revisiting the initial question of handling ‘big data’ posed in the introduction, what comes to the fore is, on the contrary, the potential of digital methods for handling a large corpus of serial sources.

Apart from this, the modeling of my data also forced me to actively engage with my source material in a specific way. Adapting my perception of

---

24 McCarty, Willard, What’s going on?, in: *Literary and Linguistic Computing* 23 (2008), 254.

the sources to the requirements of data modeling meant to not look at them as individual disputes but rather to search for features they shared. Consequently, trials came to the fore as situations in which people and institutions interact with each other under relatively fixed rules. Bringing those configurations into focus also encompasses a particular attention for deviations one may not have seen otherwise. This tension brought to awareness the crucial question of how one conceptualizes the legal field and thereby became a reminder of one's own spatial and temporal embeddedness, one's 'Standortgebundenheit': As a twenty-first-century historian one tends to view the jurisdiction as a field thoroughly regulated by the state. But in eighteenth-century Pondicherry it was, in fact, a highly dynamic field where the trading company admittedly claimed jurisdictional power but rules for the everyday co-existence of the different groups were nonetheless subject to negotiations and were constantly on the move.

Moreover, this thinking in terms of abstract patterns and structures is a way to engage with the material that differs from the more usual perspective of a historian. As a hermeneutically working scholar one rather uses the material as a set of sources providing information about different disputes between historical actors which may give some indication of more general proceedings and, in a synopsis, of historical shifts. Appropriating the sources to a database's specifications, on the contrary, entails the necessity to develop *one* grid that can be applied to *all* sources. This means the historian from the outset has to radically detach her attention from the individual case and to redirect it to a level that as a rule encompasses more than one case and is oriented towards a formal description of all possible varieties.

In the end, it is exactly this change of perspective and the encounter of – from a historian's view – unaccustomed ways of looking at material that brings to awareness one's own premises and that renders the modeling of data enriching even if the actual data is not analyzed.

## Bibliography

*Anonymus*, Arrêt du 31.03.1739, in: Gnagou Diagou (ed.), Arrêts du Conseil Supérieur de Pondichéry, Pondicherry: Société de l'histoire de l'Inde française, 1935, vol. 1, 58–59.



- Anonymus*, Arrêt du 20.03.1747, in: Gnagou Diagou (ed.), Arrêts du Conseil Supérieur de Pondichéry, Pondicherry: Société de l'histoire de l'Inde française, 1935, vol. 1, 178–181.
- Anonymus*, Procès-verbal du 22.06.1718, in: Edmond Gaudart (ed.), Procès-verbaux des délibérations du Conseil Supérieur de Pondichéry, Pondicherry: Société de l'histoire de l'Inde française, 1913, vol. 1, 191–192.
- Anonymus*, Tarabellion c/veuve Luc Carachi, 09.12.1766, in: Jean-Claude Bonnan (ed.), Jugements du Tribunal de la Chaudrie de Pondichéry. 1766–1816, 2 vols., Pondicherry: Institut français de Pondichéry, 1999, vol. 1, 3.
- Bonnan, Jean-Claude* (ed.), Jugements du Tribunal de la Chaudrie de Pondichéry. 1766–1816, 2 vols., Pondicherry: Institut français de Pondichéry, 1999.
- Brauner, Christina/Flüchter, Antje*, Introduction: The Dimensions of Transcultural Statehood, in: Christina Brauner/Antje Flüchter (eds.), The Dimensions of Transcultural Statehood, Leipzig: Leipziger Universitätsverlag, 2015, 7–26.
- Dirks, Nicholas B.*, Castes of Mind: Colonialism and the Making of Modern India, Princeton, New Jersey: 2001.
- Dönecke, Anna*, 'Le chapeau ou la toque': Rechtliche Vielfalt und soziale Diversität in Pondichéry im 18. Jahrhundert, in: Christina Brauner/Antje Flüchter (eds.), Recht, Ordnung, Diversität, Bielefeld: transcript, forthcoming.
- Epple, Angelika/Erhart, Walter* (eds.), Die Welt beobachten: Praktiken des Vergleichens, Frankfurt a. M.: Campus, 2015.
- Flanders, Julia/Jannidis, Fotis*, Data Modeling, in: John Unsworth/Raymond George Siemens/Susan Schreibman (eds.), A New Companion to Digital Humanities, Chichester, West Sussex, UK/Malden, MA, USA: Blackwell, 2016, 229–237.
- Flüchter, Antje*, Structures on the Move: Appropriating Technologies of Governance in a Transcultural Encounter, in: Antje Flüchter/Susan Richter (eds.), Structures on the Move: Technologies of Governance in Transcultural Encounter, Berlin/London: Springer, 2012, 1–27.
- Houllemare, Marie*, La justice française à Pondichéry au XVIIIe siècle, une justice en 'zone de contact', in: Éric Wenzel/Éric de Mari (eds.), Adapter le droit et rendre la justice aux colonies: Thémis outre-mer, XVIe-XIXe siècle, Dijon: Éditions Universitaires de Dijon, 2015, 147–157.

- Jannidis, Fotis*, Grundlagen der Datenmodellierung, in: Fotis Jannidis/Hubertus Kohle/Malte Rehbein (eds.), *Digital Humanities: Eine Einführung*, Stuttgart: Metzler, 2017, 99–108.
- McCarty, Willard*, What's going on?, in: *Literary and Linguistic Computing* 23 (2008), 253–261.
- McCarty, Willard*, *Humanities Computing*, Basingstoke: Palgrave Macmillan, 2014.
- Menski, Werner*, Jean-Claude Bonnan, Jugements du tribunal de la chaudrie de Pondichéry 1766–1817, in: *Indo-Iranian Journal* 46 (2003), 369–371.
- Parasher, Gauri*, Between Sari and Skirt: Legal Transculturality in Eighteenth-Century Pondicherry, in: Christina Brauner/Antje Flüchter (eds.), *The Dimensions of Transcultural Statehood*, Leipzig: Leipziger Universitätsverlag, 2015, 56–77.
- Pillai, Ananda Ranga*, *The Private Diary of Ananda Ranga Pillai: Dubash to Joseph François Dupleix [...] A Record of Matters Political, Historical, Social, and Personal from 1736 to 1761*, ed. and transl. by J. Frederick Price, Madras: Government Press, 1904.
- Quamen, Harvey/Bath, John*, Databases, in: Constance Crompton/Richard J. Lane/Ray Siemens (eds.), *Doing Digital Humanities: Practice, Training, Research*, London/New York: Routledge, 2016.
- Schmale, Wolfgang*, Big Data in den historischen Kulturwissenschaften, in: Wolfgang Schmale (ed.), *Digital Humanities: Praktiken der Digitalisierung, der Dissemination und der Selbstreflexivität*, Stuttgart: Franz Steiner, 2015.



# Looking for Textual Evidence

## Digital Humanities, Middling-Class Morality, and the Eighteenth-Century English Novel

---

*Ralf Schneider, Marcus Hartner, Anne Lappert*

### 1. Introduction

In our contribution to this edited volume we present a discussion of an attempt to identify and locate literary manifestations of the idea of the “virtuous social middle” in a large corpus of eighteenth-century English novels with the help of methods and tools from Digital Humanities (DH).<sup>1</sup> This attempt was situated within the larger context of a research project on comparative practices in the eighteenth-century novel as part of the Collaborative Research Center (CRC) 1288 “Practices of Comparing” funded by the German Research Foundation (DFG). Our project started from three assumptions. The first was the traditional assumption held in literary history about the close connection between socio-historical developments and the “rise of the novel”<sup>2</sup> from “the status of a parvenu in the literary genres to a place of dominance” during the eighteenth century.<sup>3</sup> Second, we assumed that the cultural construction of the “the middle order of mankind”<sup>4</sup> and its concomitant claims about a supposedly heightened sense of ‘middle-class’ morality

---

1 *Wahrman, Dror*, *Imagining the Middle Class: The Political Representation of Class in Britain, c. 1780–1840*, Cambridge: Cambridge University Press, 1995, 64.

2 *Watt, Ian*, *The Rise of the Novel: Studies in Defoe, Richardson, and Fielding*, Berkeley: University of California Press, 1957.

3 *Rogers, Pat*, *Social Structure, Class, and Gender, 1660–1770*, in: J.A. Downie (ed.), *The Oxford Handbook of the Eighteenth-Century Novel*, Oxford: Oxford University Press, 2016, 39.

4 *Goldsmith, Oliver*, *The Vicar of Wakefield*, Oxford: Oxford University Press, 2006 [1766], 87.

was accompanied by a range of social processes of comparing.<sup>5</sup> After all, constructions of social identity tend to rely heavily on processes of othering, and comparing plays a vital role in the construction of self and other. Third, we assumed that in the emerging medium of the novel in the period under investigation, concepts of middle-class social identity were negotiated through particular *literary* strategies of comparing, whose textual manifestations can be found specifically in textual representation of characters and character constellations. Ultimately, the underlying value system concerning class identity in a novel ought to manifest itself also in the way that the behavior or dispositions of characters are described and evaluated in comparison as either desirable and adequate, or as despicable and inappropriate. As part of our strategy of substantiating those three assumptions, our project aimed at providing a more extensive review of the textual representations of social virtues and vices in the eighteenth-century English novel than available in traditional scholarly accounts of the topic so far.

In order to achieve this aim, we decided to turn to the methods of DH. We planned to identify, with the help of different types of word searches (see below), recurrent expressions that refer to social behavior in either positive or negative terms. We expected a diachronic development to be visible across the corpus, e. g., similar to the way concepts of gentility changed their semantics during the period under consideration.<sup>6</sup> None of our expectations were met, however, as we will demonstrate below. This prompted a reconsideration of our search strategies and ultimately led to the insight that practices of comparing and social-identity construction may be more implicit in

---

5 In the following, we will employ the term 'middle-class' as a synonymous stylistic variation to expressions such as 'middle order', 'middle rank', the 'middling sorts', etc. We are aware that the application of the terminology of class to discussions of eighteenth-century society is contested and comes with certain conceptual problems. For introductions to the term and concept of class in early modern Britain, see *Corfield, Penelope J.*, *Class by Name and Number in Eighteenth-Century England*, in: *History* 72 (1987) and *Cannadine, David*, *Class in Britain*, London: Penguin, 2000, 27, 31.

6 The concept of the gentleman, for example, changed from the narrow denotation of a man of noble birth to the more widely applicable notion of a man displaying a set of 'genteel' (moral) qualities and behaviours. During this "social peregrination" of the term, it lost "its oldest connotations of 'gentle' birth and 'idle' living, so that, in the later eighteenth century, individual vintners, tanners, scavengers, potters, theatre managers, and professors of Divinity could all claim the status, publicly and without irony" (*P.J. Corfield*, *Class by Name and Number*, 41).

literature than in other discourses and function in different ways. In what follows, we will first sketch the socio-cultural context of our corpus, in which the novels contribute to the negotiation of middle-class morality. We will then briefly engage with the question of the applicability of DH methods in the analysis and interpretation of literature, before we document some of our text searches and discuss the results.

## 2. Inventing the superiority of the middling classes

On the opening pages of Daniel Defoe's *The Life and Adventures of Robinson Crusoe* (1719) the title character's elderly father lectures the youthful protagonist on his place in the social fabric of eighteenth-century Britain. In his attempt to dissuade the restless and adventure-seeking Robinson from "[going] abroad upon Adventures", he emphasizes his son's birth into the "the middle State" of society.<sup>7</sup> This he declares to be "the best State in the World, the most suited to human Happiness" as it is neither "exposed to the Miseries and Hardships, the Labour and Sufferings of the mechanick Part of Mankind", nor is it "embarrass'd with the Pride, Luxury, Ambition and Envy of the Upper Part of Mankind".<sup>8</sup> While those remonstrations unsurprisingly fail to convince the young Robinson Crusoe, they articulate a sentiment of 'middle-class' complacency found with increasing frequency in literary and philosophical writings over the course of the eighteenth century. Defoe's fictional character constitutes only one voice in an increasingly audible choir within the cultural discourse of the period that promotes the idea of the 'middle order' as possessing a distinct and superior quality. Though this idea was neither new nor universally acknowledged,<sup>9</sup> it became increasingly

---

7 Defoe, Daniel, Robinson Crusoe, ed. Michael Shinagel, New York: Norton, 1994 [1719], 5.

8 Ibid.

9 On competing models of the social structure of the period, such as the notion of a bipolar "crowd-gentry reciprocity" (Thompson, E. P., *Customs in Common: Studies in Traditional Popular Culture*, New York: New Press, 1993, 71) and the persistent traditional belief in a providentially ordained, universal and hierarchical order of social layers (e.g. Tillyard, E. M. W., *The Elizabethan World Picture: A Study of the Idea of Order in the Age of Shakespeare*, London: Chatto & Windus, 1967 [1942]), see the discussion in D. Cannadine, *Class in Britain*, 24–56. With regard to the notion of the superiority of the 'middle-class', see also French, who argues that the aristocracy and gentry retain their dominant economic and politi-

attractive to those who saw themselves as belonging to this particular segment of society.<sup>10</sup> Building on the notion of a “virtuous social middle”<sup>11</sup> initially developed in Aristotle’s *Politics*,<sup>12</sup> they actively engaged in the discursive construction of the middle order as a distinct social group not only by discussing its political and economic importance for the nation,<sup>13</sup> but also by emphatically emphasizing its moral value.<sup>14</sup>

David Hume, for example, thought that the upper classes were too immersed in the pursuit of pleasure to heed the voices of reason and morality, while “the Poor” found themselves entirely caught up in the daily struggle for survival.<sup>15</sup> As a result, in his view, only the “middle Station” affords

“[...] the fullest Security for Virtue; and I may also add, that it gives Opportunity for the most ample Exercise of it [...]. Those who are plac’d among the lower Rank of Men, have little Opportunity of exerting any other Virtue, besides those of Patience, Resignation, Industry and Integrity. Those who are advanc’d into the higher Stations, have full Employment for their Generosity, Humanity, Affability and Charity. When a Man lyes betwixt these two Extremes, he can exert the former Virtues towards his Superiors, and the latter towards his Inferiors. Every moral Quality, which the human Soul is susceptible of, may have its Turn and, and be called up to Action: And a Man may, after this Manner, be much more certain of his Progress in Virtue, than where his good Qualities lye dormant, and without Employment.”<sup>16</sup>

---

cal power in Britain throughout the eighteenth century and beyond (*French, Henry*, *Gentlemen: Remaking the English Ruling Class*, in: Keith Wrightson (ed.), *A Social History of England: 1500–1750*, Cambridge: Cambridge University Press, 2017, 269, 280). See also *Muldrew, Craig*, *The ‘Middling Sort’: An Emergent Cultural Identity*, in: Keith Wrightson (ed.), *A Social History of England: 1500–1750*, Cambridge: Cambridge University Press, 2017 on the emergence of the “Middling Sort” as a cultural identity during the early modern period.

10 *D. Cannadine*, *Class in Britain*, 32–33.

11 *D. Wahrman*, *Imagining*, 64.

12 *Aristotle*, *The Politics*, trans. Carnes Lord, Chicago: University of Chicago Press, 1984, IV.11.

13 *D. Cannadine*, *Class in Britain*, 42.

14 The protagonist Charles Primrose in Oliver Goldsmith’s *The Vicar of Wakefield*, for example, sees “the middle order of mankind” as the social sphere that is home to “all the arts, wisdom, and virtues of society” (87–88).

15 *Hume, David*, *Of the Middle Station of Life*, in: Thomas H. Green/Thomas H. Grose (eds.), *David Hume, The Philosophical Works*, Aalen: Scientia Verlag, 1964 [1742], 4:376.

16 *Ibid.*, 4:376–4:377.

The passage indicates that Hume sees the middling class's superior virtue as the result of a sociological process. By being exposed to a wider and more complex range of social life, individuals from the middle ranks are forced to develop greater moral sensitivity and power of judgement. While he thus attempts a philosophical explanation,<sup>17</sup> other contemporary authors champion middle-class virtue in a more simplistic fashion by rhetorically foregrounding the idea of a stark contrast between the "generous Disposition and publick Spirit" of members of the middling ranks and the "Depravity and Selfishness of those in a higher Class".<sup>18</sup>

It is important to note once more that such arguments about the (moral, economic, political, etc.) superiority of a distinct middle order or class, were less "an objective description of the social order" in Britain than "a way of constructing and proclaiming favourable ideological and sociological stereotypes" of those who found themselves hierarchically situated between the poor and the powerful.<sup>19</sup> In this context, the development of the eighteenth-century novel as a distinct literary genre on the fast-growing market for printed material can be seen instrumental in the emergence of the (self-) image of the middle class as an economically relevant and culturally powerful social group.<sup>20</sup> Written by (predominantly) middle-class authors for a (predominantly) middle-class audience,<sup>21</sup> the novel played an important role in the invention and promotion of this group's social identity, especially by con-

---

17 For a discussion of Hume's position in relation to that of Aristotle, see *Yenor, Scott*, David Hume's Humanity: The Philosophy of Common Life and Its Limits, Basingstoke: Palgrave, 2016, 114–119.

18 *Thornton, William*, The Counterpoise: Being Thoughts on a Militia and a Standing Army, London: Printed for M. Cooper, 1752. Quoted from the unpaginated preface.

19 *D. Cannadine*, Class in Britain, 32.

20 The connection between the "rise of the novel" and the emerging middle class was first discussed in *J. Watt*, Rise of the Novel, and *Habermas, Jürgen*, The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society, trans. Thomas Burger, Cambridge: Polity, 2015 [1962]. For a survey of perspectives after those authors, see *Cowan, Brian*, Making Publics and Making Novels: Post-Habermasian Perspectives, in: *J. A. Downie* (ed.), The Oxford Handbook of the Eighteenth-Century Novel, Oxford: Oxford University Press, 2016.

21 Hunter points out that the readership of the novel was never restricted to one specific group only. In contrast to the argument presented here, he holds that "the characteristic feature of novel readership was its social range [...] and the way it spanned the social classes and traditional divisions of readers" (*Hunter, Paul*), The Novel and Social/Cultural



tributing to the illustration and dissemination of the concept of middle-class morality.<sup>22</sup> As a result, a preoccupation with the figure of the individual forced to navigate morally complex situations, together with the frequent vilification of characters from aristocracy and gentry, as well as a complacent middle-class contentedness with being placed in the ‘best’ social stratum, set the tone for much eighteenth-century prose writing.<sup>23</sup> However, while the general connection between “the emergence of a bourgeois public sphere” and “the rise of novel writing and -reading” has long been treated as “a standard feature” of the period’s literary and cultural history,<sup>24</sup> the aesthetic and narratological dimensions<sup>25</sup> of the “invention” of middle-class superiority<sup>25</sup> still remain a productive field of study.

For this reason, our research project within the CRC 1288 “Practices of Comparing” set out to investigate the novel’s contribution to eighteenth-century negotiation of social identity and morality by focusing on the play of narrative and stylistic strategies that constitute an important aspect of this contribution. As we are traditionally trained literary scholars, the methodological thrust of our project lay in the informed manual analysis of an ambitious, yet manageable corpus of some twenty carefully selected novels from the period. We specifically decided to focus on classical narratological analyses of aspects such as narrative situation, focalization, and perspective structure<sup>26</sup> as well

---

History, in: John Richetti (ed.), *The Cambridge Companion to the Eighteenth-Century Novel*, Cambridge: Cambridge University Press, 1996, 19).

22 *Nünning, Vera*, From ‘honour’ to ‘honest’: The Invention of the (Superiority of) the Middling Ranks in Eighteenth Century England, in: *Journal for the Study of British Cultures* 2 (1994).

23 For more detailed surveys of the eighteenth-century novel and its contexts, see *Nünning, Ansgar*, *Der englische Roman des 18. Jahrhunderts aus kulturwissenschaftlicher Sicht. Themenselektion, Erzählformen, Romangenres und Mentalitäten*, in: *Ansgar Nünning* (ed.), *Eine andere Geschichte der englischen Literatur. Epochen, Gattungen und Teilgebiete im Überblick*, Trier: WVT, 1996, and the contributions in *Richetti, John* (ed.), *The Cambridge Companion to the Eighteenth-Century Novel*, Cambridge: Cambridge University Press, 1996 and *Downie, J. A.* (ed.), *The Oxford Handbook of the Eighteenth-Century Novel*, Oxford: Oxford University Press, 2016.

24 *P. Rogers*, *Social Structure*, 47.

25 *V. Nünning*, From ‘honour’ to ‘honest’.

26 *Fludernik, Monika*, *An Introduction to Narratology*, London: Routledge, 2009, *Wenzel, Peter* (ed.), *Einführung in die Erzähltextanalyse: Kategorien, Modelle, Probleme*, Trier: WVT, 2004, *Nünning, Ansgar*, *Grundzüge eines kommunikationstheoretischen Modells*

as on the representation of fictional characters.<sup>27</sup> Our individual (close) readings indeed produced results that hermeneutically seem to confirm our assumptions of a middling-class preoccupation with social identity. Nevertheless, we remained painfully aware of the limited scope of our project design regarding the number of texts that we were able to incorporate into our investigation. And we wondered if we could complement the traditional literary analyses of our research by turning to DH in the attempt to engage with at least some aspects of our research on a digital and somewhat broader textual basis.

### 3. Between close and distant reading: using DH methods for literary analysis and interpretation

While the tentative origins of DH reach back into the first half of the twentieth century,<sup>28</sup> most of its methods and research questions fully emerged only during the past few decades. One branch of the wider field of DH has concerned itself with literary texts; and its exploration of the relationship between literature and the computer has taken many shapes. One major issue is the production and increasing availability of electronic (and scholarly) editions of primary and secondary works. This development has significantly widened access to literary texts and now plays a vital role in the preservation of books and other textual materials;<sup>29</sup> it has forced libraries and academic institutions to develop new data policies and technological solutions for storing and providing access to primary and secondary literature. Also, not only computer-related genres such as literary hypertexts

---

der erzählerischen Vermittlung: Die Funktion der Erzählinstanz in den Romanen George Eliots, Trier: WVT, 1989.

27 *Margolin, Uri*, Character, in: David Herman (ed.), *The Cambridge Companion to Narrative*, Cambridge: Cambridge University Press, 2007, *Eder, Jens/Jannidis, Fotis/Schneider, Ralf (eds.)*, *Characters in Fictional Worlds: Understanding Imaginary Beings in Literature, Film, and Other Media*, New York: de Gruyter, 2010.

28 *Thaller, Manfred*, *Geschichte der Digital Humanities*, in: Fotis Jannidis/Hubertus Kohler/Malte Rehbein (eds.), *Digital Humanities. Eine Einführung*, Stuttgart: J. B. Metzler, 2017, 3–4.

29 *Shillingsburg, Peter L.*, *From Gutenberg to Google: Electronic Representations of Literary Texts*, Cambridge: Cambridge University Press, 2006.

have emerged,<sup>30</sup> but digital technologies have dramatically changed both the publishing and book industry, so that many literary and scholarly texts nowadays are read not from the printed page but from the displays of e-book devices.

In the context of those developments, the impact of the digital revolution on the academic infrastructure of the humanities is without question. But while computers have long found their place even in the offices of the most technophobic academics, and while even the rear-guard of traditional literary scholars use digital information retrieval systems such as electronic library catalogues and databanks, there is still widespread resistance to some other applications of digital methods in literary research. And indeed, in the realm of literary analysis and interpretation things look a bit complicated. On the one hand, textual analysis can very well apply digitized methods, in ways comparable to the strategies of computational and corpus linguistics. In the wide field of stylometrics, for instance, large corpora of texts can be scanned for the co-occurrence of particular textual features, which can then help trace historical developments in literary language, attribute authorship, or define genres.<sup>31</sup> Also, the themes that dominate a text can be extracted by topic modeling.<sup>32</sup> On the other hand, when it comes to the *interpretation* of literary works, there is some skepticism as to the ability of computer programs to support human readers in tasks of that complexity. Although textual analysis is always the *basis* for interpretation, interpretation is usually performed, after all, by highly educated, well-informed academic readers with a hermeneutic interest in exploring the meaning – or meanings – of a text. The main interest in interpretation lies in investigating a text's combi-

---

30 Ryan, Marie-Laure, *Avatars of Story*, Minneapolis: University of Minneapolis Press, 2006, Ensslin, Astrid, Hypertextuality, in: Marie-Laure Ryan/Lori Emerson/Benjamin J. Robertson (eds.), *The Johns Hopkins Guide to Digital Media*, Baltimore: Johns Hopkins University Press, 2014.

31 Burrows, John, *Delta: A Measure for Stylistic Difference and a Guide to Likely Authorship*, in: *Literary and Linguistic Computing* 17 (2002), Jannidis, Fotis/Lauer, Gerhard, *Burrows's Delta and Its Use in German Literary History*, in: Matt Erlin/Lynne Tatlock (eds.), *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, Rochester: Camden House, 2014. See also the extensive introduction and survey by Juola, Patrick, *Authorship Attribution*, in: *Foundations and Trends in Information Retrieval* 3 (2006), 233–334, <http://dx.doi.org/10.1561/1500000005>.

32 Jannidis, Fotis, *Quantitative Analyse literarischer Texte am Beispiel des Topic Modelings*, in: *Der Deutschunterricht* 5 (2016).

nation of thematic, aesthetic and rhetorical features which are understood to be culturally embedded in complex ways. Both ample contextual research and the close scrutiny of textual features are therefore generally considered prerequisites of literary interpretation.

The 'distant' reading, i. e., the computerized analysis of textual patterns in texts, that DH have introduced to literary scholarships, thus looks fairly incompatible at first sight with the close reading and interpretation strategies practiced by the scholar trained in literary hermeneutics. Franco Moretti famously spoke of distant reading as "a little pact with the devil: we know how to read texts, now let's learn how not to read them".<sup>33</sup> But the advantage of distant reading is that it allows scholars to detect features across a number of texts that could only with difficulty and considerable use of resources be tackled by individual close readings. While the computer may lack the ability to detect 'qualitative' differences, it is its promise of a seemingly boundless quantitative analytical scope that turns it into a potentially powerful analytic tool. Moreover, DH not only offers the opportunity to extend existing research strategies in a quantitative fashion, but the playful exploration of digital tools may also lead to unexpected results and even contribute to the emergence of new research strategies. Emphasizing the productive power of playfulness and creativity, Stephen Ramsay advocates an informal "Hermeneutics of Screwing Around" as a valid computer-based research strategy for the Digital Age in an influential paper.<sup>34</sup> Concerned with the limited scope of the hermeneutical (close) readings in our project, we were intrigued both by this lure of quantitative analysis and the emergence of the "somewhat informal branch of text interpretation delightfully termed *screwmeneutics*" after Ramsay.<sup>35</sup> Therefore, we decided to embark on a complimentary investigation of the textual manifestations of some concepts of middle-class virtue in the eighteenth-century novel with the help of DH.

---

33 Moretti, *Franco*, *Distant Reading*, London: Verso, 2013, 48.

34 Ramsay, *Stephen*, *The Hermeneutics of Screwing Around; or What You Do with a Million Books*, in: Kevin Kee (ed.), *Pastplay: Teaching and Learning History with Technology*, Ann Arbor: University of Michigan Press, 2014 [2010].

35 McCurdy, *Nina et al.*, *Poemage: Visualizing the Sonic Topology of a Poem*, in: *IEEE Transactions on Visualization and Computer Graphics* 22 (2016), 447.

#### 4. From search to research: some examples

Our approach to using DH was unusual in so far as we did not take the more common route from distant to close reading but proceeded vice versa. Since we had already invested considerable effort in the (close) reading and analysis of our original corpus of ca. twenty eighteenth-century novels, we began our journey into the field of DH equipped with a solid set of expectations about the literary negotiation of social identity during the period under investigation. Starting from the hermeneutical findings of our investigation, we then attempted to corroborate our results, by taking our research into the realm of computing, more precisely, by expanding the corpus of novels under investigation and developing ideas on how DH tools could help us to support our arguments. Our first step in this process was to expand our text base by creating a digital corpus of 55 novels (see the list in the appendix to this article), thus more than doubling the number of texts. We decided to look at some of the most well-known novels from the eighteenth century as well as to include some lesser known works that were however well received during the period in question. Further, we intentionally included works from different genres such as sentimental novels, gothic novels, coming-of-age stories and adventure novels, in order to do some justice to the considerable variety and diversity in eighteenth-century literary production.<sup>36</sup>

Already during the process of compiling and preparing the corpus, however, we encountered the first methodological challenges. While DH offers a great variety of tools and approaches, digitized texts are only ever suitable for a research purpose as they are prepared accordingly. In other words, if we were to look for complex sentence structures, or even narrative patterns conveying middle-class ideology, these structures would have to be tagged beforehand in each text. This means that passages that we consider as good examples for such patterns would have to be identified and electronically annotated accordingly in the hidden plane of text information, the markup. Not only did we need digital copies of all novels, but a lot of tagging by hand would have been necessary. The reason is that no program can automatically

---

36 See J. Richetti, *The Cambridge Companion, Nünning, Ansgar and Vera, Englische Literatur des 18. Jahrhunderts*, Stuttgart: Klett, 1998, and Bakscheider, Paula R./Ingrassia, Catherine (eds.), *A Companion to the Eighteenth-Century English Novel and Culture*, Chichester: Wiley-Blackwell, 2006.

mark up more complex structural features such as comparisons between characters that are not made explicit on the textual level, but are evoked through characters acting differently in comparable situations, a strategy frequently used in prose fiction. To tag the texts for such features would be a very time-consuming process that presupposes an answer to our original question, namely what role practices of comparing play on a structural level in the textual constructions of social identity. This question would need to be answered before the markup could begin, since these structures would have to be analyzed before they could then be tagged in all texts of our corpus. We would further risk to exacerbate the danger of confirmation bias that is structurally inherent to our approach anyway, as we would run the danger of finding exactly what we placed there during the tagging process. The sheer number of working hours that would have to be put into creating new digital versions with tags made *this* type of digital research impractical for a first, tentative and playful digital exploration of our expanded corpus of eighteenth-century novels.

As a consequence of these first challenges we moved away from the idea to investigate complex syntactic and narrative structures, and turned to word and phrase searches as a feasible alternative, for which an array of DH tools are available, and for which simple text files suffice.<sup>37</sup> In this context, our assumption was that key terms denoting middle-class virtues and vices would be detectable in abundance across the novels of our corpus. While programs such as *AntConc* are especially promising when zooming in on individual texts, *Voyant* proved to be more efficient when searching larger collections of texts. Generally speaking, it is interesting to look at the frequency of words within one text and within a corpus, since words that occur very frequently (except for function words such as conjunctions or articles, which we excluded from all searches) are likely to hint at the thematic focus of a text. Sometimes, however, the opposite of an expected word frequency may be revealing, too, as was the case in the searches we document below. Since we were also interested in diachronic developments, we began by using *Voyant*, which offered a direct comparison of word frequencies and the context

---

37 Project Gutenberg is the most easily accessible online text collection for such purposes. Although random checks of Project Gutenberg texts against the printed scholarly editions we had read suggested that the former are not always entirely reliable, we decided that for the first stage of word searches, the results were unlikely to be heavily distorted.

of their appearance across the corpus as a whole. We added the year of publication to the title in order to have the novels appear in chronological order of their publication, so that any diachronic changes would be immediately visible. Since the larger framework of our project was the study of the forms and functions of practices of comparing, our very first tentative approach was to run searches for words and particles that explicitly produce comparisons (such as *more/less than*, and words containing comparatives or superlatives ending on *-er* and *-est*). The result was that comparative words and particles occurred indeed frequently in our corpus (“more”=14196 times, “less”=2531, “than”=12161, “like”=4555). However, looking closer at our results it became apparent that words such as *more* were not always used to create an explicit comparison, but in many cases appeared in other contexts, such as to emphasize the expressed meaning (‘still the more’), or to indicate temporality (‘once more’) in phrases like ‘little more than’, ‘still the more’, ‘many more’ and ‘once more’ (see fig. 1). Hence, the results of the context search put the result of the word frequency in question and provided a first indication that comparing in prose fiction might work in less explicit ways than in some other discourses.

Fig. 1: Word search for “more” and immediate contexts

Document	Left	Term	Right
15) 174...	to pay us such interest: I thought what the interest would come to,' with much	more	of the same kind; but I have, I believe, satisfied you with this taste. "Ily
15) 174...	to support one who kept pace with the expenses of Sir George Gresham. "It is	more	than possible that the distress I was now in for money, and the impracticability of
15) 174...	subject of my serious deliberation; and I had certainly resolved on it, had not a	more	shameful, though perhaps less sinful, thought expelled it from my head."—Here he hesitated a
15) 174...	life began to be numbered among my wants; and what made my case still the	more	grievous was, that my paramour, of whom I was now grown immoderately fond, shared the
15) 174...	fear of seeing his ghost." "I shall shortly doubt, Partridge," says Jones, "whether thou art	more	brave or wise."—"You may laugh at me, sir, if you please," answered Partridge; "but
15) 174...	never found a horse in my life: but I'll tell thee what, friend, thou wast	more	lucky than thou didst know of, for thou didst not only find a horse, but
15) 174...	acquainted him that he had been misinformed as to the sum taken, which was little	more	than a fifth part of what he had mentioned. "I am sorry for it with
15) 174...	whether he had most feared my death or wished it, since he had so many	more	dreadful apprehensions for me. At last, he said, a neighbouring gentleman, who had just recovered
15) 174...	he partly owed his preservation to my humanity, with which he professed himself to be	more	delighted than he should have been with my filial piety, if I had known that
15) 174...	desires of a foolish old fellow. Such solicitations, however, had no effect, and I once	more	saw my own home. My father now greatly solicited me to think of marriage; but
15) 174...	the necessaries of life, I betook myself once again to study; and that with a	more	inordinate application than I had ever done formerly; The books which now employed my time
15) 174...	the Holy Scriptures; for they impart to us the knowledge and assurance of things much	more	worthy our attention than all which this world can offer to our acceptance; of things
15) 174...	to think all the time I had spent with the best heathen writers was little	more	than labour lost: for, however pleasant and delightful their lessons may be, or however adequate
15) 174...	were the worst of company to each other: but what made our living together still	more	disagreeable, was the little harmony which could subsist between the few who resorted to me
15) 174...	and promised to bring him the rest next morning; and after giving him a little	more	advice, took my leave. "I was indeed better than my word; for I returned to
15) 174...	that side." "This apothecary was one of the greatest politicians of his time. He was	more	delighted with the most poultry packet, than with the best patient, and the highest joy
15) 174...	I had no arms, to have executed vengeance on his baseness. "I was now once	more	at liberty; and immediately withdrawing from the highway into the fields, I travelled on, scarce
15) 174...	an end to all my apprehensions of danger, and gave me an opportunity of once	more	visiting my own home, and of enquiring a little into my affairs, which I soon
15) 174...	to be the ringleader. Thus, as our duty to the king can never be called	more	than our second duty, he had discharged us from this by making it incompatible with
15) 174...	of rebellion in any people." "I promise you, sir," says Jones, "all these facts, and	more	, I have read in history, but I will tell you a fact which is not
15) 174...	this kingdom in favour of the son of that very King James, a professed papist,	more	bigoted, if possible, than his father, and this carried on by Protestants against a king

We then turned to other word searches. Collecting results from our (close) reading of the selected text from our original corpus and in the playful spirit of “screwmenetics”<sup>38</sup> we developed a list of terms that describe behavior and

38 N. McCurdy et al., Poemage, 447.

dispositions in negative and positive ways that we considered to be important for the negotiation of social identity in eighteenth-century English novels. In particular, we decided to look for positively and negatively connotated adjectives, but also noun phrases used in characterization by narrators and other characters, or in self-characterization. With this we aimed to make apparent the contrast between what were considered desirable or undesirable character traits and actions and how these conceptions changed throughout eighteenth-century literature. For this purpose, we created two lists of adjectives we came across in our close reading process and in our reading of secondary literature on the construction of social identity in the eighteenth century.<sup>39</sup> In the group of positive terms, we had collected such words as “gentle”, “gallantry” and “virtuous”; the negative ones included “foppish”, “conceited”, “impertinence”, etc. We then added other terms from these and related semantic fields and complemented the adjectives and adverbs with the pertinent noun phrases in an attempt not to overlook relevant textual manifestations. This gave us a list that included the words “gentle” and “gentleman”, “gallant” and “gallantry”, “grace”, “graceful”, “gracious” and “graciousness”, “polite” and “politeness” “virtuous” and “virtue” for the positively connotated behaviors and attitudes; the negatively connotated ones included “fool” and “foolish”, “fop”, “foppish” and “foppery”, “disagreeable”, “conceited” and “conceitedness”, “vulgar” and “vulgarity”, “impertinent” and “impertinence”, “impetuous” and “impetuosity”, as well as “negligent” and “negligence”. For efficient text searching, the truncated forms of these words were used.<sup>40</sup> Our list then had for instance *gentle\**, *tender\**, *grac\**, *gallant\**, *polit\**, *sweet\** *virtu\**, *modest\**, *moderat\**, on the positive side, and *fop\**, *fool\**, *disagreeab\**, *conceited\**, *vulgar\**, *impertinen\**, *impetuo\**, *negligen\** on the negative.

Figure 1 shows the frequency of both the negative and the positive search terms across our corpus. Since the corpus was organized chronologically, the graphic ought to show whether certain terms were used more or less frequently in later publications than in earlier ones. As we see in the diagram, usage did vary considerably, but this variation shows no indication

---

39 V. Nünning, From ‘honour’ to ‘honest’, D. Wahrman, *Imagining*.

40 Using truncated forms, i. e., a word stem closed by an asterisk, allows the system to find instances of the stem in all variations and word classes; for example, *gentl\** would not only include the results for “gentle”, but also for “gently”, “gentleman”, “gentlemanly”, etc.



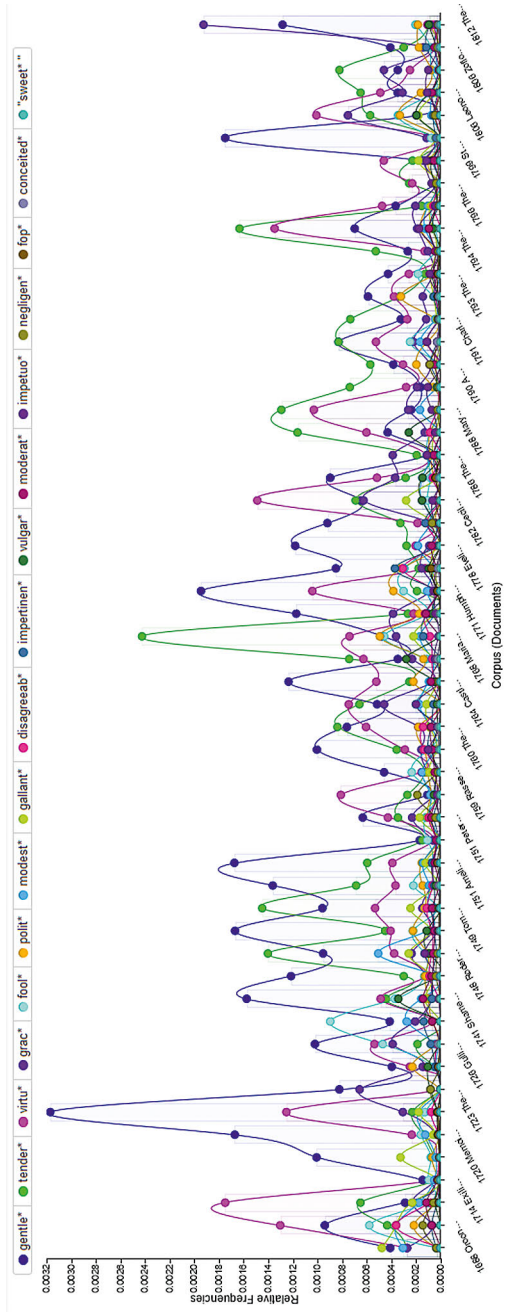
of being related to diachronic changes during the time period. Frequencies rather vary from text to text. In fact, while individual texts may deviate from the median in a significant fashion, the overall frequency of the terms under investigation seems to remain more or less consistent over the entire eighteenth century as far as our corpus is concerned.

The underlying assumption guiding our approach was that the social changes in the understanding of the virtues and vices listed above would somehow be reflected by changing word frequencies. Especially for *gentle*\* did we expect to find a significant diachronic development, as notions of gentility changed from a rather narrow denotation of gentle birth to an understanding of polite behavior by the end of the century that made it possible for men from a significantly wider range of society to claim the status of a “Gentleman” (see FN 6). Contrary to our expectations, however, we were unable to discern significant developments in our search result.

While *gentle*\* indicated at least a slight discernible decrease of usage (see fig. 2), none of the other terms offered a visible indication of a diachronic development. Put differently, word *frequencies* did not hint at the emerging construction of a middle-class identity during the period, as described by eighteenth-century social history. One possible explanation for this may be that frequency cannot capture what a term *means*: While narrators and characters in late eighteenth-century novels may use all variations of the words “gentle”, “gentleman”, etc. as frequently as those in the early phase, they may simply mean different things by those terms.

With this possible explanation in mind, we decided to turn away from questions of diachronic development within the eighteenth century. Our next step was to look at the total word frequency of our search terms in the entire corpus. In order to corroborate our assumption that these terms play a significant role in the topics of the novels, we checked their position in the list of the most frequently appearing words within the body of novels under consideration. However, we were once more disappointed. The word count showed that out of the words we were looking for, most were situated in the lower ranks of the count, whereas words such as ‘said’, ‘Mr’, ‘time’ and ‘little’ came up top of the list (see fig. 3). From our search list, only *gentleman* managed to enter the top 100 at position 81, followed by *virtue* at 239. The results for our negative terms proved to be even less impressive with, for instance, *fool*, reaching only the top 2000 of the most frequently used words in the corpus. Our positive terms generally ranked higher than our negative

Fig. 2: Frequency of negative and positive terms across the corpus



terms with *virtue* at position 239 (1370 occurrences), *agreeable* at 440 (892 occurrences), *sweet* at 450 (881 occurrences), and *tender* at 299 (1163 occurrences). None of the negative terms made it above *fool* at position 1420 and with 329 occurrences. With all our negative terms ranking rather low and quite a number of our positive terms ranking comparatively higher and with a look at the most frequent words (especially “dear”, “great”, and “good”), one may speculate whether character traits might have been negotiated more in terms of stating an ideal during the period. This would mean that texts rather state what should be aimed for, while at the same time only implicitly hinting at negative traits and behaviors and hence, at what to avoid. On the other hand, our experience with words and particles that explicitly produce comparison showed that word frequency tells us little about the contexts of use, and hence little about the diverse meanings individual words can take on in different contexts. Such a bold claim would therefore need more data via context searches or a more elaborate analysis via close reading.

Fig. 3: Top 40 most frequent words in the corpus

	Term	Count	Trend
1	said	202...	
2	mr	107...	
3	time	101...	
4	little	8293	
5	man	8070	
6	good	7974	
7	great	7865	
8	sir	7596	
9	shall	7385	
10	know	7325	
11	lady	6809	
12	mrs	6299	
13	make	6287	
14	heart	6073	
15	miss	5835	
16	think	5762	
17	having	5610	
18	thought	5470	
19	mind	5372	

	Term	Count	Trend
20	came	5237	
21	say	5072	
22	house	4990	
23	day	4978	
24	till	4869	
25	dear	4837	
26	cried	4834	
27	lord	4831	
28	life	4796	
29	father	4689	
30	told	4565	
31	come	4563	
32	like	4555	
33	love	4515	
34	way	4473	
35	soon	4392	
36	long	4389	
37	young	4310	
38	hand	4206	
39	let	4141	
40	world	4027	

Fig. 4: Position of the word “conduct” in the word count

179	light	1680	
180	cause	1675	
181	order	1669	
182	true	1669	
183	ill	1651	
184	castle	1649	
185	affection	1645	
186	heaven	1645	
187	conduct	1639	
188	god	1622	
189	pleased	1617	
190	countenance	1613	
191	matter	1608	
192	certain	1606	

After none of our search terms had turned out to feature prominently among the most frequent words in the corpus, our next step was to turn to what we *could* find on the list of most frequent words (figs. 3 and 4). For this we went through this list looking for terms we felt to exhibit some kind of relationship to contemporary discussions of social identity. In this way, we found that comparably frequently used in our corpus were “honour” (no. 54 in the word-frequency list), “poor” (no. 47), “character” (no. 141) and “conduct” (no. 187; see fig. 4), with the word “conduct”, referring to the overall comportment of a person. From those results, we considered “conduct” to be particularly interesting. The term, appearing most frequently in Wollstonecraft’s *Maria, Or, The Wrongs of Woman* (1798) and least frequently in Fielding’s *Shamela* (1741), is not only eponymous to the eighteenth-century genres of the conduct book and the conduct novel, but generally constitutes a key concept of the literary and cultural movement of sensibility.<sup>41,42</sup> For this reason, we decided to play around some more and searched for the word “conduct” in the sentimental novels of our corpus separately.<sup>43</sup> Once more, we received a fairly inconclusive diagram (fig. 5): Between the middle and the end of the eighteenth century, sentimental novels feature the term “conduct” in varying ways.

While interesting for the formulation of new research questions,<sup>44</sup> this did not help us in terms of our thesis on the literary negotiations of social identity. In fact, the visualization suggested that a diachronic change in the

---

41 V. Nünning, *From 'honour' to 'honest'*.

42 The low result for *Shamela* could be interpreted in different ways. On the one hand, it could mean that this text, being a parody of one of the most influential of the early sentimental novels, wanted to avoid the term by way of taking a critical stance on the genre of the sentimental novel, which was heavily influenced by the conduct book. On the other hand, Fielding may simply have counted on the reader to realise that both the original and the parody deal with conduct, without having to make that explicit.

43 The sentimental novels or parodies thereof in our corpus are, in chronological order: Samuel Richardson’s *Pamela* (1740), Henry Fielding’s *Shamela* (1741) and *Amelia* (1751), Laurence Sterne’s *Tristram Shandy* (1759), Oliver Golding’s *The Vicar of Wakefield* (1766), Henry Mackenzie’s *The Man of Feeling* (1771), Tobias Smollet’s *Humphrey Clinker* (1771), Frances Burney’s *Evelina* (1778), Maria Edgeworth’s *Castle Rackrent* (1800) and Jane Austen’s *Sense and Sensibility* (1811).

44 Such as: Is a separation of a genre and its parodies necessary, and if so, how can such a distinction be upheld? In how far does the illustration present a visualization of genre negotiations by means of comparison? Is this wave movement even coincidental due to the novels in the corpus?

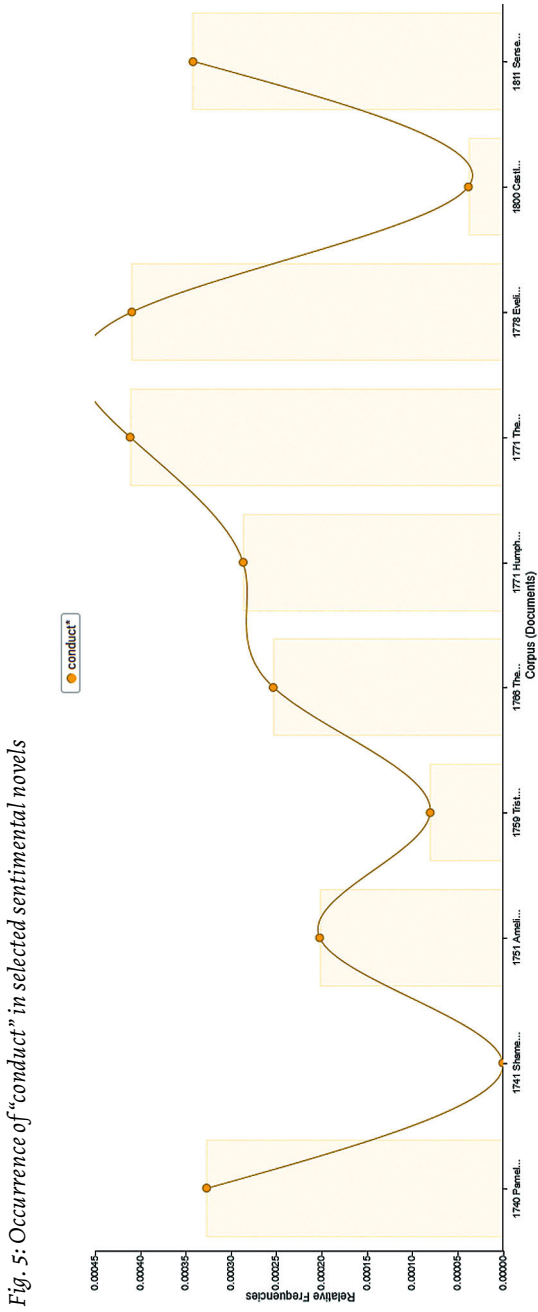


Fig. 5: Occurrence of “conduct” in selected sentimental novels

usage of particular words is rather difficult to argue for, based on the type of distant reading we engaged in our work with Voyant. Another visualization tool by Voyant offers users the opportunity to look for the context and the co-occurrence of individual terms in a corpus. Here it became apparent that “conduct”, while mainly appearing as a noun in connection with adjectives that qualify it, also appears as a verb, and does so most frequently in our Gothic novels (figs. 6 and 7). With their tendency to set the action in regions both temporally and spatially remote from eighteenth-century England, the Gothic novels can comment on contemporary English society at best by implication, so that the latter finding pointed once more at the need for further close reading and interpretation.

Fig. 6: Examples for sentences containing “conduct” (1)

1) 1688 ...	a man who has led them on to battle with	conduct	and success; of whom I shall have occasion to speak
1) 1688 ...	to arms, and the occasions given him, with the good	conduct	of the old general, he became, at the age of
1) 1688 ...	moans with sighs and tears. What reports of the prince's	conduct	were made to the king, he thought good to justify
1) 1688 ...	forgetting how time ran on, and that the dawn must	conduct	him far away from his only happiness, they heard a
1) 1688 ...	and to the world, that depended on his courage and	conduct	. But he made no other reply to all their supplications
1) 1688 ...	word and honour he would find the means to re-	conduct	him to his own country again; assuring him, he had
2) 1713 ...	Aunt, my Lady Martial, a virtuous Matron, under whose pr...	conduct	I might learn a little of the Town Politeness, its
2) 1713 ...	in which I had greatest Need of her Counsel and	conduct	; and as most young People have too great an Opinion
2) 1713 ...	try if Jealousy would work upon me, but all my	conduct	had been with Caution and Circumspection, quite different from Passion
3) 1714 ...	said Clelia, that I always liv'd at Rome, under the	conduct	of my wise and honourable Parents, the noble Fabius, my
3) 1714 ...	if I may so say) into the right Way, and	conduct	us thro' unknown Paths, to what we desire, or, at
3) 1714 ...	my Duty. I who had, by my disobedient and unwary	conduct	, in some Degree tarnish'd the Glory of my illustrious Family
3) 1714 ...	that even the Line of Reason is not able to	conduct	me through its wild Mazes. On every Hand I see
3) 1714 ...	Accommodation, after so many great and dangerous Fati...	conduct	you to your Apartment. Book 3 Having left Clarinthia and

Fig. 7: Examples for sentences containing “conduct” (2)

31) 177...	is lawfully entitled. It is true, that Mrs. Mirvan would	conduct	this affair with more delicacy than Mrs. Selwyn; yet, perhaps
31) 177...	when, with the eye of penitence, thou reviewest thy past	conduct	I Hear, then, the solemn, the last address, with which the
31) 177...	so obviously, without considering the strange appearance ...	conduct	. Alas, my dearest Sir, that my reflections should always be
31) 177...	you are, you will adopt a very different style and	conduct	in future." I then rose, and was going, but he
31) 177...	in the world who would have any influence over my	conduct	." "And will you, then, restore to me that share of
31) 177...	in great indignation; and assuring him I would make his	conduct	as public as it was infamous-I left the house
31) 177...	manner almost unanswerable, besought me to leave to hi...	conduct	of the affair, by consenting to be his before an
31) 177...	very sorry for it!-Lord Orville must himself think this	conduct	strangely precipitate." "No, my dear, you are mistaken; Lord Orville
31) 177...	seemed something so little-minded in this sudden change of	conduct	, that, from an involuntary motion of contempt, I thanked her
31) 177...	will certainly be offended; but if you allow me to	conduct	you, though she may give the freer scope to her
31) 177...	My sole view is to explain the motive of my	conduct	in a particular instance, and to obviate the accusation of
31) 177...	may know I dare defend, as well as excuse, my	conduct	." CLEMENT WILLOUGHBY." What a strange letter! how proud and how
31) 177...	for no more; the chaise now waits which is to	conduct	me to dear Berry Hill, and to the arms of
32) 177...	told him that a servant from the Baron waited to	conduct	him to the Castle. He took leave of Wyatt's wife

## 5. Discussion

The results of our investigations with Voyant were unexpected to say the least. They are not only at odds with important voices in secondary literature,<sup>45</sup> but they also contradict our own close reading experiences that confirm the conceptual relevance of the listed virtues and vices in the portrayal of characters in the eighteenth-century novel in general. Our expectation was to find diachronic developments of the words used to describe presumably middle-class virtues and flaws displaying an increase of frequency towards the end of the eighteenth century. We based our expectations on the assumption that the social identity of the ‘middling’ classes began to be constructed in negotiations in and beyond literature during this time period.<sup>46</sup> By use of visualization tools we expected to be able to localize the moment these negotiations entered literature on a word level, but instead the results indicate that as far as our searched terms and our corpus are concerned no such change is traceable. Confronted with these findings, we naturally began to question our search strategies, including the list of terms we had thought to be so prominent in eighteenth-century discussions of virtues and vices. But we also wondered whether the infrequent appearance of those terms and the lack of clearly discernible diachronic developments in their application could also be explained differently, for example, by considering the traditional distinction made in literary studies between *telling* and *showing*.<sup>47</sup> Thus, we speculated that our findings may indicate a tendency to show virtues by means of the description of behaviors rather than by naming them explicitly. However, such a claim can only be upheld by a closer analysis in terms of close reading as a complementary method to the usage of DH tools.

Further, it seemed that when working with computation techniques, there is the danger that significant differences between texts belonging to the various subgenres of the novel that constitute the overall corpus may disappear from view. While a scholar has certain background information on literary and cultural history available in close reading, a computer is rather

---

45 E. g., A. Nünning, *Der englische Roman*.

46 D. Wahrman, *Imagining*, Schwarz, L. D., *Social Class and Social Geography: The Middle Classes in London at the End of the Eighteenth Century*, in: *Social History* 7 (1982).

47 Herman, David, *Story Logic: Problems and Possibilities of Narratology*, Lincoln: University of Nebraska Press, 2002, 171–172.



ignorant towards contextual details in its application of distant reading on a text. This bears problems as well as promises. But we wondered whether these subgenre specific groups such as Gothic novels, or Sentimental novels, had not better be analyzed by searching them separately. The justification for dealing with these separate groups of novels belonging to different subgenres separately lies in literary-historical conventions and definitions of, e. g., the Sentimental Novel, or Gothic Novel. The very fact that DH overlooks such conventions and definitions in the production of data, makes us aware of their potential relevance for analysis and interpretation. In the words of McCarty, DH forces us to “ask in the context of computing what can (and must) be known of our artifacts, how we know what we know about them and how new knowledge is made”.<sup>48</sup> Just as in the case of words and particles which explicitly produce comparisons, and with the different rankings of positive and negative terms, our usage of DH tools challenged us to acknowledge that computing can only ever give us information on texts in form of data. How we read and interpret these numbers and results foregrounds the responsibility of informed research. It is easy to quickly jump to false conclusions if the numbers seem to support the desired argument. But especially when we combine traditional research with DH methodology taking into consideration all the different aspects that influence the results (e. g., the corpus, the genre, the scope of each text, the relation to other literary works of the same time period, etc.) becomes a difficult yet, important task for every scholar in the humanities.

For our usage of Voyant this meant we had to realize that even when we received results that seemed to corroborate our assumptions, this did not really mean direct support for our argument in terms of numbers. It only meant that we needed to question these results again in order to avoid running the risk of prematurely interpreting unanticipated quantitative data in the light of our underlying argument. In the case of the term *conduct*, for example, we seemed to have found a frequently used word that could support our argument of the negotiation of social identity in terms of morals in the eighteen-century English novel. Instead, further testing via other visualization tools offered by Voyant made clear that this seemingly simple link between word frequency and research question offered a false security

---

48 McCarty, Willard, *Encyclopedia of Library and Information Science*, New York: Dekker, 2003, 1231.

(figs. 2, 4, and 5). Looking at the context of the usage of the word “conduct”, we could not support our argument but had to face that the various different contexts of occurrences of “conduct” varied significantly in meaning. This means that only in a few cases of the many occurrences did “conduct” actually appear in contexts that we had in mind and that supported our argument (figs. 7 and 8). Another example for the need to treat numeric results with caution was the result of the search term *fool*. The 1741 novel *Shamela* by Henry Fielding, which was written as a parody of Samuel Richardson’s highly influential sentimental novel *Pamela*, showed a peak in the frequency of the word “fool” (551 occurrences) in comparison to the other novels. Strikingly, the second highest frequency of the word “fool” was actually found in Richardson’s *Pamela*, with 175 occurrences. The temptation to construct some intertextual correlation between both texts with regard to their top positions in the word count for “fool” was great: Fielding might have picked up an inherently significant feature of Richardson’s novel and exaggerated that for the purposes of satire. However, when we took into account the overall length of the two texts, this argument collapsed: While 551 occurrences of *fool* seem noteworthy in the relatively short novel *Shamela* (14.456 words), there is nothing significant about the term’s appearance in *Pamela* given the total length of this work. With 227.407 words Richardson’s novel is over fifteen times longer than that of Fielding. Thus, given the massive text of *Pamela*, the count of 175 occurrences of *fool* dwindles into comparative insignificance.

We were left with the paradox that while being considered more precise and accurate in terms of quantitative and statistical occurrences than traditional methods of close, DH actually seemed to blur any assumption of a precise answer to questions of literary analysis. In our case, DH appeared to be more suitable for finding new questions than to offer or support conclusive answers to interpretative assumptions. Voyant was able to give us the exact number of word frequencies, to tell us which word appeared how often in which novel, and even offered us to compare these frequencies across the corpus directly, while allowing us at the same time to look for the specific contexts of the words. All of this was very helpful, but mainly to question our own approach and its underlying categories. We set out to look for literary negotiations of social identity and how these were influenced by practices of comparison, just to be faced with the problem that comparison was already included in every aspect of our own approach. Instead of making clear distinctions more apparent, DH made us question these distinctions from the

start. If this were the end of it, we would come out of this experiment quite disillusioned. Instead we are inspired by what seems to offer a new methodology for approaching literary texts. While the usage of computing in literary studies is often feared to turn literary analysis into a mere equation whose solution would render all further examination of a text vain and shallow, the opposite seems to be true. DH offers a chance to engage in a more playful, more open-minded yet at the same time equally critical approach to literature and its study that eventually draws research back to the text and the question of how texts are embedded in various discourses.

## 6. Conclusion

At first sight, our engagement with text search and visualization tools for the analysis of a corpus of eighteenth-century English novels could be summarized in terms of discouragement and frustration – an experience that appears to be shared by scholars in other DH projects but that is apparently rarely admitted in DH. According to Jasmine Kirby “[w]e don’t talk enough about failure in the digital humanities”.<sup>49</sup> Our failure to corroborate some of our assumptions with numerical data, and the necessity to proceed from the observation of word counts to the wider contexts of our findings in fact triggered two insights. First, if the actions and dispositions of humans in social interaction that the eighteenth-century novel negotiates as desirable or undesirable are much less explicitly mentioned than expected, the novel must have other ways of presenting them. Second, the practices of comparing, too, appear to be situated on *other* levels than that of the text surface, at least in the corpus under scrutiny in our project. The lack of simple numerical proof garnered from distant reading was, in our case, a productive ‘failure’, because it helped us formulate the hypothesis that literary practices of comparing involve the structural juxtaposition of characters in comparable settings and plot segments. As Nina McCurdy and her colleagues have demonstrated, there is some irony in the fact that the more precision a DH tool offers, the more it makes sense to ‘screw around’ with it to render new interesting and exciting research questions. Narrow research questions in DH often offer a

---

49 Kirby, *Jasmine S.*, How NOT to Create a Digital Media Scholarship Platform: The History of the Sophie 2.0 Project, in: *IASSIST Quarterly* 42 (2019), <https://doi.org/10.29173/iq926>.

variety of open, inconclusive results, while ‘screwing around’ seems to lead to unexpected, innovative questions.<sup>50</sup> None of this narrows literary research down to a question of software engineering and mathematical bean counting, but rather computation techniques in form of tools offer a playful exchange between the traditionally trained scholar and DH to find ever new ways of reading texts together in the midst of the “*beautiful mess*” that is literature.<sup>51</sup>

What our search for textual evidence also appeared to show was that available strategies of tagging the words and passages of a text – the production of markup – could much profit from taking into account the research questions of literary scholarship. Existing markup algorithms performed autonomously by computer programs, may be helpful and time-saving, and they certainly have improved much in recent years; still, they rarely capture any of the more content-related questions pertaining to literary analysis, let alone interpretation. What, in the case of our project, really would have helped would have been the automatic isolation and tagging of passages that contain comparisons; this however, is nowhere in sight. We also encountered problems in the visualization of results, even though our corpus was, in DH terms, very small. How could meaningful illustrations be produced if hundreds, or even thousands, of books were subjected to data-mining? Visualization tools will also have to be further developed to match the research designs of the humanities better.

After our venture into DH, we still believe that no computer can ‘find out’ anything about the meaning of a text on its own. Therefore, while the scholar’s limitations are quantitative, those of computer programs appear to lie in the quality of their findings. Nor will a text be ‘readable’ to a computer at all, if it has not been previously read, processed and digitized by humans, increasingly automatized programs of parsing and tagging notwithstanding. The solution to the apparent incompatibility of close and distant reading lies, unsurprisingly, in the fact that the two strategies can, and ought to be, regarded as complementary rather than competitive, as Stephen Ramsay, among others, has argued.<sup>52</sup> As we have shown, to make use of DH methods can help literary scholars to focus and re-formulate their questions and research strategies, and to reconsider their assumptions about what literary texts do and how they do it.

---

50 N. McCurdy et al., Poemage, 447.

51 Ibid., 445.

52 S. Ramsay, The Hermeneutics.

## Appendix 1: Extended corpus of English eighteenth-century novels

1688	<i>Oroonoko</i>	Aphra Behn
1713	<i>The Amours of Bosvil and Galesia</i>	Jane Barker
1714	<i>Exilius</i>	Jane Barker
1719	<i>Robinson Crusoe</i>	Daniel Defoe
1720	<i>Memoirs of a Cavalier</i>	Daniel Defoe
1722	<i>Moll Flanders</i>	Daniel Defoe
1723	<i>The Lining of the Patch Work Screen</i>	Jane Barker
1724	<i>John Sheppard</i>	Daniel Defoe
1726	<i>Gulliver's Travels</i>	Jonathan Swift
1740	<i>Pamela</i>	Samuel Richardson
1741	<i>Shamela</i>	Henry Fielding
1743	<i>Jonathan Wild</i>	Henry Fielding
1748	<i>Roderick Random</i>	Tobias Smollett
1749	<i>Fanny Hill</i>	John Cleland
1749	<i>Tom Jones</i>	Henry Fielding
1750	<i>Harriot Stuart</i>	Charlotte Lennox
1751	<i>Amelia</i>	Henry Fielding
1751	<i>Betsy Thoughtless</i>	Eliza Fowler Haywood
1751	<i>Peter Wilkins</i>	Robert Paltock
1752	<i>The Female Quixote</i>	Charlotte Lennox
1759	<i>Rasselas</i>	Samuel Johnson
1759	<i>Tristram Shandy</i>	Laurence Sterne
1760	<i>The Adventures of Sir Launcelot Greaves</i>	Tobias Smollett
1762	<i>Millenium Hall</i>	Sarah Scott
1764	<i>Castle of Otranto</i>	Horace Walpole
1766	<i>The Vicar of Wakefield</i>	Oliver Goldsmith
1768	<i>Maria; Or, The Wrongs of Woman</i>	Mary Wollstonecraft
1769	<i>Emily Montague</i>	Frances Brooke
1771	<i>Humphrey Clinker</i>	Tobias Smollett
1771	<i>The Man of Feeling</i>	Henry Mackenzie
1778	<i>Evelina</i>	Frances Burney
1778	<i>The Old English Baron</i>	Clara Reeve
1782	<i>Cecilia</i>	Fanny Burney
1784	<i>Imogen</i>	William Godwin
1786	<i>The Heroine</i>	Eaton Stannard Barrett

1786	<i>Vathek – An Arabic Tale</i>	William Beckford
1788	<i>Mary – A Fiction</i>	Mary Wollstonecraft
1789	<i>Castles of Athlin and Dunbayne</i>	Ann Radcliffe
1790	<i>A Sicilian Romance</i>	Ann Radcliffe
1791	<i>A Simple Story</i>	Elizabeth Inchbald
1791	<i>Charlotte Temple</i>	Susanna Rowson
1791	<i>Romance of the Forest</i>	Ann Radcliffe
1793	<i>The Castle of Wolfenbach</i>	Eliza Parsons
1794	<i>Caleb Williams</i>	William Godwin
1794	<i>The Mysteries of Udolpho</i>	Ann Radcliffe
1796	<i>Memoirs of Emma Courtney</i>	Mary Hays
1796	<i>The Monk</i>	Matthew Lewis
1798	<i>Wieland</i>	Charles Brockden Brown
1799	<i>St Leon</i>	William Godwin
1800	<i>Castle Rackrent</i>	Maria Edgeworth
1806	<i>Leonora</i>	Maria Edgeworth
1806	<i>Wild Irish Girl</i>	Sydney Owenson
1806	<i>Zofloya</i>	Charlotte Dacre
1811	<i>Sense and Sensibility</i>	Jane Austen
1812	<i>The Absentee</i>	Maria Edgeworth

## Bibliography

- Aristotle, *The Politics*, trans. Carnes Lord, Chicago: University of Chicago Press, 1984.
- Backscheider, Paula R./Ingrassia, Catherine (eds.), *A Companion to the Eighteenth-Century English Novel and Culture*, Chichester: Wiley-Blackwell, 2006.
- Burrows, John, Delta: A Measure for Stylistic Difference and a Guide to Likely Authorship, in: *Literary and Linguistic Computing* 17 (2002), 267–287.
- Cannadine, David, *Class in Britain*, London: Penguin, 2000.
- Corfield, Penelope J., Class by Name and Number in Eighteenth-Century England, in: *History* 72 (1987), 38–61.
- Cowan, Brian, Making Publics and Making Novels: Post-Habermasian Perspectives, in: J. A. Downie (ed.), *The Oxford Handbook of the Eighteenth-Century Novel*, Oxford: Oxford University Press, 2016, 55–70.

- Defoe, Daniel*, *Robinson Crusoe*, ed. Michael Shinagel, New York: Norton, 1994 [1719].
- Downie, J. A.* (ed.), *The Oxford Handbook of the Eighteenth-Century Novel*, Oxford: Oxford University Press, 2016.
- Eder, Jens/Jannidis, Fotis/Schneider, Ralf* (eds.), *Characters in Fictional Worlds: Understanding Imaginary Beings in Literature, Film, and Other Media*, New York: de Gruyter, 2010.
- Ensslin, Astrid*, *Hypertextuality*, in: Marie-Laure Ryan/Lori Emerson/Benjamin J. Robertson (eds.), *The Johns Hopkins Guide to Digital Media*, Baltimore: Johns Hopkins University Press, 2014, 258–265.
- Fludernik, Monika*, *An Introduction to Narratology*, London: Routledge, 2009.
- French, Henry*, *Gentlemen: Remaking the English Ruling Class*, in: Keith Wrightson (ed.), *A Social History of England: 1500–1750*, Cambridge: Cambridge University Press, 2017, 269–89.
- Goldsmith, Oliver*, *The Vicar of Wakefield*, Oxford: Oxford University Press, 2006 [1766].
- Habermas, Jürgen*, *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*, trans. Thomas Burger, Cambridge: Polity, 2015 [1962].
- Herman, David*, *Story Logic: Problems and Possibilities of Narratology*, Lincoln: University of Nebraska Press, 2002.
- Hume, David*, *Of the Middle Station of Life*, in: Thomas H. Green/Thomas H. Grose (eds.), *David Hume, The Philosophical Works*, Aalen: Scientia Verlag, 1964 [1742], 375–380.
- Hunter, Paul J.*, *The Novel and Social/Cultural History*, in: John Richetti (ed.), *The Cambridge Companion to the Eighteenth-Century Novel*, Cambridge: Cambridge University Press, 1996, 9–40.
- Jannidis, Fotis*, *Quantitative Analyse literarischer Texte am Beispiel des Topic Modelings*, in: *Der Deutschunterricht* 5 (2016), 24–35.
- Jannidis, Fotis/Lauer, Gerhard*, *Burrows's Delta and Its Use in German Literary History*, in: Matt Erlin/Lynne Tatlock (eds.), *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, Rochester: Camden House, 2014, 29–54.
- Juola, Patrick*, *Authorship Attribution*, in: *Foundations and Trends in Information Retrieval* 3 (2006), 233–334, <http://dx.doi.org/10.1561/1500000005>.

- Kirby, Jasmine S.*, How NOT to Create a Digital Media Scholarship Platform: The History of the Sophie 2.0 Project, in: IASSIST Quarterly 42 (2019), <https://doi.org/10.29173/iq926>.
- Margolin, Uri*, Character, in: David Herman (ed.), *The Cambridge Companion to Narrative*, Cambridge: Cambridge University Press, 2007, 66–79.
- McCarty, Willard*, *Encyclopedia of Library and Information Science*, New York: Dekker, 2003.
- McCurdy, Nina et al.*, Poemage: Visualizing the Sonic Topology of a Poem, in: IEEE Transactions on Visualization and Computer Graphics 22 (2016), 439–448.
- Moretti, Franco*, *Distant Reading*, London: Verso, 2013.
- Muldrew, Craig*, The ‘Middling Sort’: An Emergent Cultural Identity, in: Keith Wrightson (ed.), *A Social History of England: 1500–1750*, Cambridge: Cambridge University Press, 2017, 290–309.
- Nünning, Ansgar and Vera*, *Englische Literatur des 18. Jahrhunderts*, Stuttgart: Klett, 1998.
- Nünning, Ansgar*, Der englische Roman des 18. Jahrhunderts aus kulturwissenschaftlicher Sicht. Themenselektion, Erzählformen, Romangenres und Mentalitäten, in: Ansgar Nünning (ed.), *Eine andere Geschichte der englischen Literatur. Epochen, Gattungen und Teilgebiete im Überblick*, Trier: WVT, 1996, 77–106.
- Nünning, Ansgar*, Grundzüge eines kommunikationstheoretischen Modells der erzählerischen Vermittlung: Die Funktion der Erzählinstanz in den Romanen George Eliots, Trier: WVT, 1989.
- Nünning, Vera*, From ‘honour’ to ‘honest’. The Invention of the (Superiority of) the Middling Ranks in Eighteenth Century England, in: *Journal for the Study of British Cultures* 2 (1994), 19–41.
- Ramsay, Stephen*, The Hermeneutics of Screwing Around; or What You Do with a Million Books, in: Kevin Kee (ed.), *Pastplay: Teaching and Learning History with Technology*, Ann Arbor: University of Michigan Press, 2014 [2010], 111–120.
- Richetti, John* (ed.), *The Cambridge Companion to the Eighteenth-Century Novel*, Cambridge: Cambridge University Press, 1996.
- Rogers, Pat*, Social Structure, Class, and Gender, 1660–1770, in: J. A. Downie (ed.), *The Oxford Handbook of the Eighteenth-Century Novel*, Oxford: Oxford University Press, 2016, 39–54.



- Ryan, *Marie-Laure*, *Avatars of Story*, Minneapolis: University of Minneapolis Press, 2006.
- Schwarz, *L. D.*, *Social Class and Social Geography: The Middle Classes in London at the End of the Eighteenth Century*, in: *Social History* 7 (1982), 167–185.
- Shillingsburg, *Peter L.*, *From Gutenberg to Google: Electronic Representations of Literary Texts*, Cambridge: Cambridge University Press, 2006.
- Thaller, *Manfred*, *Geschichte der Digital Humanities*, in: Fotis Jannidis/Hubertus Kohle/Malte Rehbein (eds.), *Digital Humanities. Eine Einführung*, Stuttgart: J. B. Metzler, 2017, 3–12.
- Thompson, *E. P.*, *Customs in Common: Studies in Traditional Popular Culture*, New York: New Press, 1993.
- Thornton, *William*, *The Counterpoise: Being Thoughts on a Militia and a Standing Army*, London: Printed for M. Cooper, 1752.
- Tillyard, *E. M. W.*, *The Elizabethan World Picture: A Study of the Idea of Order in the Age of Shakespeare*, London: Chatto & Windus, 1967 [1942].
- Wahrman, *Dror*, *Imagining the Middle Class: The Political Representation of Class in Britain, c. 1780–1840*, Cambridge: Cambridge University Press, 1995.
- Watt, *Ian*, *The Rise of the Novel: Studies in Defoe, Richardson, and Fielding*, Berkeley: University of California Press, 1957.
- Wenzel, *Peter* (ed.), *Einführung in die Erzähltextanalyse: Kategorien, Modelle, Probleme*, Trier: WVT, 2004.
- Yenor, *Scott*, *David Hume's Humanity: The Philosophy of Common Life and Its Limits*, Basingstoke: Palgrave, 2016.

# The Historical Semantics of Temporal Comparisons Through the Lens of Digital Humanities

## Promises and Pitfalls

---

*Michael Götzelmann, Kirill Postoutenko, Olga Sabelfeld, Willibald Steinmetz*

### 1. Introduction<sup>1</sup>

Until recently inquiries into the historical semantics of ‘comparison’ have mostly been limited to explicit reflections on the use and misuse, or the pertinence and impertinence, of comparisons in various historical, scientific, or political settings. The preferred object of study has been what one might call ‘comparatism’, i. e., the more or less conscious application and framing of comparison as a tool in all branches of human knowledge, and it is noteworthy that the early modern period has been very much at the center of attention in this kind of research.<sup>2</sup> From the point of view of historical semantics such an approach would correspond to a strategy that is primarily interested in the changing meanings of particular words or terms that denote, or directly refer to, comparison. Among these, the noun ‘comparison’ itself and its equivalents in other languages are of course to be considered

---

1 This article has been devised – and revised – jointly by all four authors. More specifically, Kirill Postoutenko has been responsible for section 2, Olga Sabelfeld for section 3, Michael Götzelmann for section 4, and Willibald Steinmetz for the introductory and concluding remarks (sections 1 and 5).

2 See, for example, *Richter, Melvin*, “That vast Tribe of Ideas”. Competing Concepts and Practices of Comparison in the Political and Social Thought of Eighteenth-Century Europe, in: *Archiv für Begriffsgeschichte* 44 (2002), 199–219; *Eggers, Michael*, Vergleichendes Erkennen: Zur Wissenschaftsgeschichte und Epistemologie des Vergleichs und zur Genealogie der Komparatistik, Heidelberg: Universitätsverlag Winter, 2016; *Grafton, Anthony*, Comparisons Compared: A Study in the Early Modern Roots of Cultural History, in: Renaud Gagné/Simon Goldhill/Geoffrey E. R. Lloyd (eds.), *Regimes of Comparatism: Frameworks of Comparison in History, Religion and Anthropology*, Leiden/Boston: Brill, 2019, 18–48.

in the first place, to which may be added their derivatives (verbs, adjectives, adverbs, etc.) and synonyms, as well as explicit denials of comparability. In this article, for reasons of convenience, we speak of the ‘comparison vocabulary’ or of ‘terms denoting comparison’ when we refer to the semantic field of ‘comparison’ in that narrow sense. Strangely enough, given the importance of comparison as a method in the sciences and humanities, and more broadly as a social practice in all walks of life, a traditional history of the concept ‘comparison’ and adjacent concepts in major European languages, let alone non-European languages, is still only in its very first stages.<sup>3</sup> It is obvious that the massive increase of digitized source materials in recent years, even if only searchable in a very basic fashion, has made it much easier to carry out empirically based quantitative and qualitative inquiries into the changing conjunctures, meanings and uses of the comparison vocabulary.

However, the historical semantics of comparison as we understand it in this article moves a step further. While it includes the study of the terms expressly denoting the activity of ‘comparison’ and their changing meanings over time, it also takes account of the fact that most comparisons are ‘performed’ in language without making any use of what we call the comparison vocabulary. Everyone would recognize simple sentences such as ‘*x* is like *y*’, or ‘*j* is better than *k*’, or ‘*t* is unlike *t*<sub>1</sub> in this respect, but the same as *t*<sub>1</sub> in all other respects’ as comparisons, yet, in the vast majority of cases, utterances like these are used in a rather inconspicuous manner in the course of ordinary speech. Most of these ‘comparison-performing utterances’, as we will call them here for convenience’s sake, are used in a routinized fashion without the speakers deliberately intending, or overtly communicating, that they are actually about to make a comparison. Besides studying the comparison vocabulary proper, our more ambitious aim in this article is to prepare the field for identifying the ways in which comparisons are routinely performed through speech acts such as the above-mentioned, by making use of the tools of digital semantics. This requires an extension of historical-semantic

---

3 For a recent attempt with further references to relevant literature see: *Steinmetz, Willibald*, “Vergleich” – eine begriffsgeschichtliche Skizze, in Angelika Epple/Walter Erhart (eds.), *Die Welt beobachten: Praktiken des Vergleichens*, Frankfurt/New York: Campus Verlag, 2015, 85–134; see also: *Steinmetz, Willibald*, Introduction: Concepts and Practices of Comparison in Modern History, in: Willibald Steinmetz (ed.), *The Force of Comparison: A New Perspective on Modern European History and the Contemporary World*, Oxford/New York: Berghahn Books, 2019, 1–32.

study into a field that has traditionally belonged to the domain of rhetorics, or of philosophical and linguistic attempts at systematizing speech acts, in this case comparative speech acts.<sup>4</sup>

In view of this volume's general aim of exploring ways of using digital tools for the study of comparison we assume that comparison-performing speech acts may be classified so as to produce a manageable number of formalized sequences, or patterns, that can be recognized by 'machines', i. e., specifically tagged digital corpora. The task in front of us is a multiple translation work. In order to use digital tools, we need to translate conventional comparative utterances into a series of standardized sequences (section 2), which will then have to be translated into the specific codes used for queries in digitized and previously tagged corpora (section 3), or into self-defined queries that conform to the codes of freely available taggers applicable to self-defined and purposely digitized text corpora (section 4).

In order to narrow down the scope of our explorative analysis, we have chosen to focus on one particular type of comparison only: temporal comparisons, i. e., comparisons between two or more historical points in time, or between one or more objects over the course of time. The multiple translation work involved is still complicated enough but seems reasonable compared to an attempt at translating the totality of all possible comparisons into the codes suited for digital queries. Another limitation is that we have chosen to study one language only, Modern English. The following sections present our collective work in three steps. Section 2 elaborates on the difficulties and possibilities of systematizing comparison-performing utterances in general, and temporal comparisons in particular. These reflections result in a basic typology of sentences performing temporal comparisons – a typology that subsequently helps to formulate digital queries in existing or self-defined text corpora. Section 3 uses an existing digitized and already tagged corpus, the historical corpus of British parliamentary debates from 1803 to 2005 ('Hansard Corpus'), to explore various ways of identifying temporal comparisons in that corpus and of interpreting the search results in quantitative and qualitative terms. Section 4, by contrast, describes the steps that are needed

---

4 For interesting typological reflections, but also historical cases studies on comparison-performing speech acts see *Mauz, Andreas/Sass, Hartmut von (eds.), Hermeneutik des Vergleichs. Strukturen, Anwendungen und Grenzen komparativer Verfahren*, Würzburg: Königshausen & Neumann, 2011.

when a researcher wants to work on a corpus of his or her own choice, a corpus that has not been previously digitized and tagged, in this case a series of utopias and dystopias ranging from Thomas More (1516) to contemporary examples of that literary genre. The section is in itself designed as a competitive comparison between ‘man’ and ‘machine’, i. e., between traditional and digital research methods in historical semantics.

## 2. Typologies of comparative utterances: from abstraction to practice

Before going into the details of a computer-based examination of comparative utterances, it is worth explaining – in a top-down fashion – the general logic behind our analytical procedures. As soon as scholars begin approaching comparative processes in earnest, they are confronted with a barrage of quotes from renowned thinkers of all stripes extolling the virtue and ubiquity of comparison.<sup>5</sup> Indeed, if we accept the standard approach to comparison as a cross-evaluation of several objects (*comparata*) on the basis of a more or less stable combination of criteria (*tertia*), we almost immediately drown in the sea of introspections, observations and empirical measurements undertaken by most of the living beings on our planet most of the time.<sup>6</sup> Fortunately, the focus of our project is not *comparisons at large*, but rather the *practice of comparison*, understood as an intersubjective activity occurring in the social world.<sup>7</sup> Such practices offer no direct access to the mental or perceptual activities that inform comparisons, but present their results in communicative acts that are understandable by others. Thus, when Don John in Shakespeare’s *Much Ado About Nothing* (1612) says about his brother,

---

5 See one of the earliest examples: “in omni ratiocinatione per comparationem tantum veritatem praecise cognoscamus”. Rene Descartes, “Regulae ad directionem ingenii (1619)”. *Descartes, Rene, Œuvres*. T. X, Paris: Cerf, 1908, 439.

6 See, for instance: Cheah, Pheng, The Material World of Comparison, in: *New Literary History* 40 (2009), 524. Mignolo, Walter D., Who Is Comparing What and Why, in: Rita Felski/Susan Stanford Friedman (eds.), *Comparison: Theories, Approaches, Uses*, Baltimore: John Hopkins University Press, 2013, 99.

7 See: Grave, Johannes, Vergleichen als Praxis. Vorüberlegungen zu einer praxistheoretisch orientierten Untersuchung von Vergleichen, in: Angelika Epple/Walter Erhart (eds.), *Die Welt beobachten: Praktiken des Vergleichens*, Frankfurt a. M.: Campus, 2015, 136–137.

the Prince of Aragon: “I had rather be a canker in a hedge, than a rose in his grace”,<sup>8</sup> we may be unsure about the exact *tertia* lurking in Don John’s (or Shakespeare’s) mind, but the alleged qualitative difference between Aragon’s hypothetical states – *\*being a canker in a hedge* and *\*being a rose in his grace* – are expressed clearly and unambiguously for every competent user of a natural language. In social terms, saying *N would rather [do/feel] X than Y* is practicing comparison, and the systemic nature of language allows us to hope that all statements that conform to this syntactic pattern are comparisons. A number of lexemes (such as the adjectives *similar/different*) and morphemes (the suffix *-er* in comparative adjectives *bigger/smaller* etc.) reinforce this impression of regularity: the form that practices of comparison may take in language is therefore not arbitrary but constrained by the rules and conventions of specific languages. Since grammarians and semioticians know the rules of many, if not most, verbal and visual languages, one could hope that practices of comparison in society may ultimately be reduced to a string of formalized patterns which computers might cherry-pick for us in various kinds of data.

Such hopes are not entirely unfounded: particularly among diehard representatives of linguistic structuralism the notion of language as a set of binary oppositions has always been popular. Roman Jakobson used an example from Lewis Carroll’s *Alice in Wonderland* to present this hypothesis to the general audience: “‘Did you say pig or fig?’ said the Cat. ‘I said pig,’ replied Alice.”<sup>9</sup> Jakobson commented upon this dialogue: “In this peculiar utterance the feline addressee is attempting to recapture a linguistic choice made by his addressor. In the common code of the Cat and Alice, i. e., in spoken English, the difference between a stop and a continuant consonant, other things being equal, may change the meaning of the message”<sup>10</sup> So, if the semantics of natural languages can possibly be exhaustively described by means of privative oppositions, why not try the same thing with comparisons?

---

8 Shakespeare, William, *Much Ado about Nothing* [1623], Philadelphia, PA: J.P. Lippincott Company, 2001, 53.

9 Carroll, Lewis, *Alice’s Adventures in Wonderland* [1865], Los Angeles: Enhanced Media, 2016, 32.

10 Jakobson, Roman, *The Cardinal Dichotomy of Language*, in: Ruth Nanda Anshen (ed.), *Language: An Enquiry into its Meaning and Function*, Port Washington and London: Kennikat Press, 1971, 157.

At a first glance, such a typology, if rigorously executed, works quite well. Thus, on a basic level, one might distinguish between grading comparisons, equations, differentiations, statements about comparisons, denials of comparability, etc.; and one might then go on to introduce more fine-grained distinctions between various syntactic or semantic patterns that express these and other basic types.<sup>11</sup> Indeed, it seems likely that such a classification of comparative utterances could be arranged into a set of neat privative oppositions (*a* vs.  $\neg a$ ) translatable into computer-friendly formal language without much hassle. A case in point is the difference between comparative (*a* is more/less *n* than *b*) and superlatives (*a* is the most/least *n* in the group [*b*, *c*, *d* ... etc.]). Consider the following utterances by Bossuet and Voltaire seeking to grasp the historical destiny of the Jewish people in the light of the schism between Judaism and Christianity:

“The Jews are more demolished than their temple, or city.”<sup>12</sup>

“It is certain that the Jewish nation is the most singular that the world has ever seen.”<sup>13</sup>

While Bossuet, a French conservative bishop and theoretician of history, takes a well-trodden path back to 70 CE, Voltaire, in the vein of Enlightenment, extends the meaning of the alleged ‘Jewish self-alienation’ beyond religion and history. Consequently, his emphasis is on the general ‘otherness’ of Jews, expressed in their unparalleled “singularity”; Bossuet, in contrast, is focused on the inner “abasement” of the Jewish people, which he judges to be more severe than the destruction of the Second Temple or the corresponding fall of Jerusalem. In grammatical terms, the difference between relative (Bossuet) and absolute (Voltaire) grading comparisons is crystal-clear. The first category is articulated by a combination of the adjective *more* (*plus*) placed just

11 For an elaboration on this see: *Postoutenko, Kirill*, Preliminary typology of comparative utterances: a tree and some binaries (in print). Inevitably, some of the examples and argumentation are borrowed from this article.

12 “Les Juifs sont plus abattus que leur temple et que leur ville” (*Bossuet, Jacques-Bénigne*, Discours sur l’Histoire universelle [1681], Paris: Garnier-Flammarion, 1966, 274).

13 “La nation juive est la plus singulière qui jamais ait été dans le monde” (*Voltaire [François-Marie Arouet]*, Des Juifs [1756], in: André Versaille (ed.), Dictionnaire de la pensée de Voltaire par lui-même, Bruxelles: Editions Complexe, 1994, 688).

before the first *comparatum* and the conjunction *than (que)* enclosed between the first and the other *comparata*. The second, in its turn, is conveyed by the combination of the same adjective *plus* with the immediately preceding definite article *le* which – together with the absence of *que* – turns the whole thing into a superlative: *plus ... que* → *le plus...* = *more* → *the most*. In some languages, the situation is even more straightforward as there is a special pair of morphemes responsible for the difference between comparative and superlative degrees in adjectives (for example, *-er/-est* in German). So, on the surface of it, there is an impeccable privative opposition between the absolute and the relative grading comparisons. On the one hand, the differentiation itself looks certain as the association of an expression with one semantic category excludes it from the other, and no third kind of grading comparisons is imaginable. On the other hand, the morphological alternation supporting the distinction seems to be so regular and semantically uniform that any differentiation problems appear highly unlikely. So, up to this point, the prospects of formalizing the typology of comparative utterances look quite rosy.

Alas, as soon as we leave the abstract, platonic meta-grammar hovering high in the sky and descend into the midst of social life, the typology of comparative utterances comes under attack from various sides, including (but not limited to) inconsistencies of verbal languages, social customs and political control. Even in the relatively simple and modest domain of grading comparisons there are myriad ways of verbalizing both absolute and relative grading comparisons which seriously compromise the formalizations attempted above. If we were fixed on the verbal meanings only, we had to conclude that Charles Perrault's mentioning of Louis XIV as "the incomparable prince" merely meant that no sensible comparative utterance with the Sun King as a *comparatum* were possible.<sup>14</sup> In fact, Perrault's statement is an absolute grading comparison echoing many similarly constructed references to Gods, royals and lovers.<sup>15</sup> What is more, in many situations one could observe the obliteration of the very difference between comparatives and

14 "Ces ouvrages divins où tout est admirable, // Sont du temps de Louis, ce prince incomparable." *Perrault, Charles*, *Parallèle des anciens et des modernes en ce qui regarde les arts et les sciences* [1687], München: Eidos, 1964, 10.

15 See: *Luhmann, Niklas*, *Liebe als Passion. Zur Kodierung der Intimität*, Frankfurt a. M.: Suhrkamp, 1982, 154. *Steinmetz, Willibald*, *Above/below, better/worse, or simply different? Metamorphoses of Social Comparison, 1600–1900*, in: Willibald Steinmetz (ed.),



superlatives, as the expressions from the one category are used for expressing its opposite. Thus, when the Baptist rebel John Bunyan speaks of Jesus's "desire that the worst of these worst should in the first place come unto him",<sup>16</sup> he allows for a grading of superlatives, which twists their meanings in a curious way: at any rate, the *worst of these worst* presupposes the difference in meaning between the two identical adjectives, with the second *worst* inevitably losing its superlative semantics. All in all, neither the grammar-based formulae of relative and absolute comparisons nor the very contradistinction between them withstand the test of social practice. Needless to say, the same problems plague many other categories of the typology such as, for example, equations.

Still, if numerous exceptions are taken note of and categorized, we may end up with a workable classification of comparative utterances in which inconsistencies, instead of being swept under the carpet, are seen as crucial markers of social, political and cultural influences upon comparative practices.<sup>17</sup> The first step towards this distant but achievable goal would be the ultimate abandonment of logical and grammatical determinism in the typology of comparative utterances in favor of an interactionist standpoint as pioneered by pragmatist philosophy: Whatever is perceived as a comparison by the reader, listener, etc. should work as such, no matter what shape it takes.<sup>18</sup> For a praxis-oriented theory, this sounds like a good – if not the only – way to go about comparisons. In reality, however, such an approach proves to be no less difficult.

Take the seemingly easy case of *equations*, sentences such as *x is like b*. Even within our eminently cooperative group of four authors we failed to achieve consensus on whether individual sentences taken from a large sample of examples preliminarily earmarked as equations would actually qualify as equations or not. Our sample of supposed equations was taken from dif-

---

The Force of Comparison: A New Perspective on Modern European History and the Contemporary World, New York and Oxford: Berghahn, 2019, 80–112.

16 Bunyan, John, The Jerusalem Sinner Saved [1688], Edinburgh: The Banner of Truth, 2005, 33.

17 For a first attempt, see: K. Postoutenko, Preliminary Typology.

18 It might be fitting here to draw upon the very beginnings of pragmatist semasiology forestalling "the practical turn": "Before I can think you to mean my world, you must affect my world." William James. The Meaning of Truth [1907]. – James, William, Pragmatism and Four Essays from 'The Meaning of Truth', New York: Meridian Books, 1955, 215.

ferent genres (utopias, political speeches, travel journals, etc.) to ensure the ubiquity of speech acts independent of genre-related languages. And yet, we could not agree whether a sentence like ‘He [Hitler] gets more and more like a Caesar’ should be considered an equation or not.<sup>19</sup> The difference of opinions on whether individual phrases such as these should be considered an equation was so wide that, for the time being, we chose to abandon the pragmatist rationalization of this particular category of comparison (equation).

In a second attempt, we chose *temporal comparisons* as a test case, i. e., sentences that set apart and compare a state ‘before’ with a state ‘after’, or a *now* and a *then*, etc. This time, the rate of agreement among our group of four authors was much higher, and eventually, we came up with a tentative praxis-based formalization of temporal comparisons. Essentially, we saw comparative utterances as being produced by a number of recurrent elements sequentially ordered in sentences, or groups thereof. Among the most frequent elements one could name:

- (1) Time references – both deictic (then, now, in old days, before, after N, today, in the future, earlier, etc.) and intrinsic (1.1.2017, after the war, XX ago, ever, some day);
- (2) Terms denoting either changes occurring over time (rise, drop, increase, decrease, no longer, no more, etc.) or their absence (remain, stand still, etc.);
- (3) Conjunctions and temporal adverbs: than, since, ever, then, as, etc.

The one unquestionably productive sequence is [Time reference / its absence] – ... – term of change / difference – ... – conjunction – ... – time reference. An example of such a sequence would be: *In 1970 there were more smokers than today*. In self-comparisons, i. e., comparisons in which not two or more different objects, but different states of the *same* object in different times are evaluated, a standard sequence would be Object N – ... terms of change or its absence – ... – Object N. An example for that kind of temporal comparison would be: *London remains London*.

These short remarks may suffice to show that temporal comparisons are relatively easy to detect, reasonably distinct and well suited to be typecast

---

19 Shire, William, Berlin Diary – The Journal of a Foreign Correspondent, 1934–1941, New York: Knopf, 1942, 7 (entry for 20 April 1937).

by means of formalized sequences. For these pragmatical reasons we chose temporal comparisons as a starting point for our inquiries into the advantages and disadvantages of digital tools for elucidating the historical semantics of comparisons more generally.

### 3. Case study I: temporal comparisons in parliamentary interaction

Parliamentary debates are an important source for an investigation of comparisons, and of temporal comparisons in particular, because of parliaments' specific features as interactional environments. First, the interaction in parliaments is characterized by *pro et contra* argumentation,<sup>20</sup> and argumentation generally, according to Chaim Perelman, can hardly avoid the use of comparisons "where several objects are considered in order to evaluate them through their relations to each other".<sup>21</sup> Secondly, parliamentary interaction is based on deliberative rhetoric that aims at decision making. Parliament offers a political arena for the negotiation of opposite points of view in a competitive mode of speaking between different parties or between opposition and government. The participants of a debate rarely accept or reject statements completely, but most of the time modify, question or add arguments. For these reasons, parliamentary debates are particularly suitable for analyzing comparison-performing speech acts as well as explicit rejections or reevaluations of comparisons, for example when speakers identify certain *tertium* or *comparata* as controversial or start questioning established relations between compared units. Another reason why parliamentary debates are an exceptional source for any inquiry into the changing uses of language is, of course, that in many countries they are now available in digitized form over long time spans. Possibly the highest standard of digitization so far has been achieved for the British parliamentary debates of the 19th and 20th centuries assembled in the Hansard Corpus.

---

20 Palonen, Kari, Concepts and Debates. Rhetorical Perspectives on Conceptual Change, in: Willibald Steinmetz/Michael Freeden/Javier Fernández Sebastián, Conceptual History in the European Space, New York and London: Berghahn, 2017, 101.

21 Perelman, Chaim/Olbrechts-Tyteca, Lucie, The New Rhetoric: A Treatise on Argumentation, trans. John Wilkinson and Purcell Weaver, 2nd ed., Notre Dame and London: University of Notre Dame Press, 1958, 242.

### 3.1 The Hansard Corpus 1803–2005: 1.6 billion words and some taggers

The Official Reports of British parliamentary debates provided by Hansard date back to the year 1803 and are available in several digitized versions.<sup>22</sup> In 2015, the *Hansard Corpus 1803–2005* (HC) was completed and put online. It contains nearly all the speeches held in both houses of the British parliament, Lords and Commons, from 1803 through to 2005.<sup>23</sup> The uniqueness of this corpus lies in the combination of a big amount of data and a semantic annotation of the entire corpus with the *Historical-Thesaurus-based Semantic Tagger* (HTST).<sup>24</sup>

The HTST provides an annotation of *all* lexical units in a text under certain grammatical and semantic criteria. The semantic classification of the HTST is based on the *Historical Thesaurus of English* (HT), a project that started in the pre-digital age (1964) and was undertaken at the University of Glasgow.<sup>25</sup> The developers of the HTST describe the method of their classification as follows:

“The semantic classification is based primarily on a systematic analysis of the content of the Oxford English Dictionary, with other content from additional dictionaries of English. To this end, words are arranged into categories by the

22 Hansard offers two versions of the *Historic Hansard 1803–2005* <https://www.hansard-corpus.org/>: The debates in xml format <http://www.hansard-archive.parliament.uk> and the data base behind the *Hansard-Corpus 1803–2005* to read online <https://api.parliament.uk/historic-hansard/index.html>. The *Hansard Online* website presented by the UK parliament offers alongside the historical debates the recent sittings of both chambers, too <https://hansard.parliament.uk/>. Unfortunately, the very new platform *Hansard at Huddersfield* could not be considered for this article <https://hansard.hud.ac.uk/site/index.php> [accessed: 13.03.2019].

23 For the project see: *University of Glasgow*, SAMUELS Project (Semantic Annotation and Mark-Up for Enhancing Lexical Searches), <https://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/samuels/> [13.03.2019].

24 Piao, Scott *et al.*, A time-sensitive historical thesaurus-based semantic tagger for deep semantic annotation, in: *Computer Speech & Language* 46 (2017), 113–135. This article gives further detailed information on the development and structure of the tagger.

25 <https://ht.ac.uk/>. The researchers in that project also developed the *Historical Thesaurus of Old English*, <https://oldenglishtesaurus.arts.gla.ac.uk/> [accessed: 13.03.2019]. Kay, Christian, Diachronic and synchronic thesauruses, in: Philip Durkin (ed.), *The Oxford Handbook of Lexicography*, Oxford: Oxford University Press, 2015, 367–380.

concepts they express, with successive subdivision of these categories delineating ever more precise sub-concepts within a concept."<sup>26</sup>

An important asset of the HT is that its taxonomy of words is 'time sensitive', i. e., that it takes account of historical changes in the meanings of words. More than other existing thesauri, the HT is therefore capable of eliminating the ambiguities of words that are due to semantic change in a short-term or *longue durée* perspective. The HTST is the first annotation tool that makes use of the time sensitive semantic classification of the HT. The classification of the HT starts at the highest level with three main categories: *the world*, *the mind*, and *society*. Each of these main categories is then subdivided into several further, ever more precise subcategories with lists of individual words assigned to each of the categories and subcategories. Depending on the historical dictionary entries, an individual word may be allocated to several subcategories or even main categories in the hierarchy.

The tagging of the Hansard Corpus (HC) is based on the hierarchy of the HT but assigns particular codes to each of the categories and subcategories. Thus, the category 'Time' (AM) is subdivided into 'Spending Time' (AM:01), 'Duration' (AM:02), 'Particular Time' (AM:03) and so on. These codes can be used to start search queries, either on their own or – what is more interesting – in combination with searches for other categories or particular words or parts of speech. For an unspecific query a large category (e. g., AM – Time) may be more useful than a smaller one (e. g., AM:04:b – a month/calendar month). Thus, the query **{AM:04:b} \* war** (month in combination with the word 'war') results in collections of sentences that contain formulations like 'months of war', 'months after war', 'in November the war would have ended by now', and so on.

The semantically tagged Hansard Corpus thus provides the opportunity to search not only for occurrences of single words, but also for syntactic and semantic patterns. Therefore, our rough typology of comparison-performing utterances (see section 2) has to be further refined and adapted for inquiries in the HC. When starting queries on temporal comparisons in the HC we need to go beyond the formalized basic types of syntactic sequences outlined above. For each lexical unit, the precise function and position in the sequence have to be clearly defined before starting a query.

---

26 S. Piao et al., *Tagger*, 115.

One of the most advantageous features of the HC is its annotation of time references. Searching for time references on a large scale may be a promising avenue to detect patterns of temporal comparison in the corpus. The HC considers as time references not only numerals referring to dates (1952), temporal adverbs (*yesterday*) or nouns (*future*) but also combinations of different parts of speech which constitute a time reference (*at the present time*). While numerals referring to dates and adverbs or nouns referring to time might perhaps be put together “manually” in a search list, it seems impossible that an individual researcher should be able to imagine all possible combinations of time references in advance and assemble them in a search list. By contrast, the HC provides this search option which will be further discussed below.

Apart from time references, another important indicator of temporal comparisons is the assertion that entities have changed across time or that a change has occurred between two points in time. This is usually expressed by using verbs or nouns denoting change, such as *rise*, *decline*, *increase*, etc. Due to the finely graduated semantic annotation of the HC we are capable of searching for those verbs or nouns without concretizing the terms beforehand. Of course, this feature applies not only to terms denoting change but also to many other semantic fields that we might be interested to study.

Fig. 1: Hansard Corpus 1803–2005, Interface

The screenshot displays the interface for the Hansard Corpus (British Parliament). At the top, it identifies the corpus as containing 7.6 million speeches and 1.6 billion words from 1803 to 2005. The user is identified as O SABELFELD, with links for history, lists, and logout.

The interface features a search sidebar on the left with the following sections:

- DISPLAY:** Options for LIST, CHART, KWIC, and COMPARE.
- SEARCH STRING:** A text input field for the search query.
- COLLOCATES:** A section for finding words that appear together.
- POS LIST:** A section for filtering by part of speech.
- SEMANTIC CATEGORIES | WORDS:** A section for semantic filtering, including buttons for RATIONCH, SEARCH, and RESET.
- SHOW DECADE | SPEAKER:** A section for filtering by decade and speaker, with dropdown menus for selection.
- IGNORE:** Two lists of years (1990, 1980, 1970, 1960, 1950) that can be added to an ignore list.
- SORTING AND LIMITS:** Options for sorting by FREQUENCY and setting a MINIMUM FREQUENCY (currently set to 20).
- CLICK TO SEE OPTIONS:** A link to expand search options.

The main area of the interface is a large grid with columns representing years from 1800 to 2000 in 10-year increments. The grid is currently empty, indicating no search results have been displayed.

At the bottom right, there is a **Personal Information** panel with the following details:

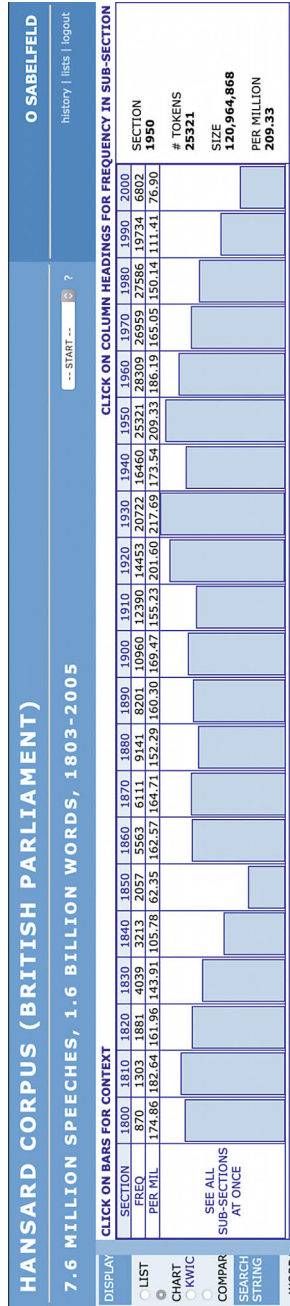
- Institution:** University Siefeld (Update)
- Country:** GERMANY (Update)
- Activity:**
  - Last login (before this time): (Utah time; same as Denver)
  - Queries (24 hours / 30 days): 0 / 92
  - Corpora used (30 days): HANS (92)
  - KWIC lists (24 hours | saved): 0/500 [0%] | 12210/15000 [81%]
  - Customized word lists: 0
- Status:**
  - Access level: 1 / 3
  - Usage limits: [CLICK TO SEE](#)

Figure 1 shows the user interface of the HC. It is divided into three parts that display different kinds of information. Whereas the left-hand side contains options to define search queries, the two boxes on the right-hand side display the search results. The box in the upper right part shows relative or absolute frequencies of the words or sequences searched for by time periods (decades, individual years), whereas the box in the lower right part shows lists of individual occurrences of the words or sequences searched for, sorted by the criteria defined on the left-hand side. Several options are available to define queries. The box labeled ‘display’ on the upper left-hand side refers to the way in which search results will appear in the boxes on the right-hand side. Four options are offered: *List*, *Chart*, *KWIC*, and *Compare*. Depending on the type of queries some of the offered four possibilities may be more useful than others. Thus, for someone interested in a diachronic conceptual history, it may be more significant to see a *chart* showing the relative and absolute frequency of the concept over decades than just a *list* with the number of occurrences per decade. *KWIC* (Keyword in Context) provides a traditional concordance view of words or phrases around the sequence. The last option *compare* allows contrasting two different words or sequences under various criteria (to be defined), for example frequency or collocations.

Figure 2 illustrates the query **compar\*** limited to the House of Commons with the display option *chart*. The columns show the number of hits and their frequency per decade. Each column (decade) contains another ten columns with detailed frequency of the lemma for every year in that decade. By choosing one of the years, a list of occurrences containing the lemma in that year appears in the box below. Lemmatization of the verb “compare” or the noun “comparison” would work for every query as well, but the short form with an asterisk “compar\*” offers the highest scope of possible parts of speech regarding this word (e. g., comparable, comparatively, etc.).

The next section on the upper left-hand side (‘search string’) is content-related, but for basic searches only the first box (*word[s]*) has to be filled. For a distinct collocation the field *collocates* has to be completed with a single word or a part of speech. By clicking on that field but leaving it empty, all the collocations will be displayed and listed whether by frequency or relevance. *Pos list* helps if one is not sure about the abbreviation for a part of speech and wants to include it in the sequence searched for. Finally, *semantic* refers to the semantic taxonomy described above and allows to search for either categories or individual words.

Fig. 2: Frequency of *compar*\* in the House of Commons (HC) 1803–2005





The next section allocates limitations like *decades*, *speaker*, *parties*, *party in power* or *chamber* (Lords or Commons). It is possible to choose only one parameter, several parameters at the same time, or none at all. In the latter case the query will run all through the whole corpus. Unfortunately, a search by subject matter of debates is not possible.

The last two sections ‘sorting and limits’ and ‘options’ improve the usability of the displayed results.

In the HC, it is possible to formulate queries that combine semantical categories, parts of speech (general and specific, e. g., general adjective, general comparative adjective, general superlative adjective, etc.), particular words and expressions, unspecified synonyms of words, and words in all possible grammatical forms. Punctuation can also be included and treated like a lexical unit in a sequence. The following chart shows some important elements of queries for temporal comparisons in the HC.

Table 1: Important elements of queries for temporal comparisons in the HC

Syntax1 <sup>a</sup>	Meaning within the query
*	Wildcard for any number of letters (even in a word, e. g. un*) but only one word in a sentence. Necessary for part of speech categories (see below).
[pos]	Part of speech in an exact grammatical form.
[pos*]	Part of speech in every grammatical form (wildcard).
?	Wildcard for one letter, can be part of a word on every position in this word.
=	Synonyms of a word. If [[=word]] HC searches for this word and its synonyms, whether noun, verb, or adjective, or other.
[]	For all forms of a word.
-	Exclusion of exact words or parts of speech.
	Separation of words or parts of speech to search for any of these words.

a For a more extensive overview see: *Alexander, Marc/Davies, Mark*, Hansard Corpus 1803–2005, [https://www.hansard-corpus.org/help/syntax\\_e.asp](https://www.hansard-corpus.org/help/syntax_e.asp) [accessed: 29.04.2019].

Semantic categories 1 <sup>b</sup>			
{AM}	Time		
{AM:04}	Period	{AM:08}	Relative time
{AM:04:a}	Year	{AM:08:b}	The Present (time)
{AM:04:a:01}	Season	{AM:08:c}	The past
{AM:04:b}	A months/calendar month	{AM:08:c:01}	Historical period
{AM:04:c}	A day/twenty-four hours	{AM:08:c:02}	Antecedence/being earlier
		{AM:08:c:03}	Oldness/ancientness
		{AM:08:d}	The future/time to come
		{AM:08:d:02}	Newness/novelty, recency

b For all semantic categories and options to search for a semantic category of an arbitrarily chosen word see *M. Alexander/M. Davies*, Hansard Corpus 1803–2005, <https://www.hansard-corpus.org/SemTags1.asp> [accessed: 29.04.2019].

Parts of speech 1 <sup>c</sup>			
[a*]	article	[nn*]	common noun, neutral for number
[csa*]	as (as conjunction)	[p*]	Pronoun
[csn*]	than (as conjunction)	[ppls2*]	3rd person plural subjective personal pronoun (they)
[d*]	demonstrative	[ppis2*]	1st person plural subjective personal pronoun (we)
[j*]	general adjective	[rgr*]	comparative degree adverb (more, less)
[jjr*]	general comparative adjective	[v*]	Verb
[jst*]	general superlative adjective	[word*],[pos*]	word as particular part of speech
[n*]	Noun	[[=word]]	word and its synonyms

c For a list of all included parts of speech that is taken from the UCREL tagger see *M. Alexander/M. Davies*, Hansard Corpus 1803–2005, <http://ucrel.lancs.ac.uk/claws7tags.html> [accessed: 29.04.2019]; Some possible combinations for queries based on parts of speech. *M. Alexander/M. Davies*, Hansard Corpus 1803–2005, [https://www.hansard-corpus.org/help/posList\\_e.asp](https://www.hansard-corpus.org/help/posList_e.asp) [accessed: 29.04.2019].

### 3.2 Six basic sequences to search for temporal comparisons

Having briefly described the HC and its capabilities, we will now introduce six formalized sequences of semantic and syntactic components that may help to detect temporal comparisons in the Hansard Corpus. The list does not claim to be exhaustive but still contains essential components that can be extended and modified for further inquiry. Also, every formalized sequence consists of several varieties of nearly identic sequences that only differ from each other in the space between the units or the specification of semantic components. The example sentences in the following paragraphs are mainly taken from parliamentary debates of the post-war periods after World War I and II.

Our first formalized sequence contains sentences using comparatives. Obviously, the core of those sentences is formed by a general comparative adjective (*better, faster, stronger*) and the conjunction *than*. These components can be combined in different ways, for example: **{AM} than \* {AM}; [jɹ\*] [csn\*][csa\*] {AM}; [n\*] [v\*] [jɹ\*] [csn\*] {AM}**. These constructions may also lead to sentences of the type *never happier than before* that are meant as superlatives but formulated as a graded comparison. That case illustrates once more that our classification is based on purely formal criteria, not on considerations related to content. Here is an example sentence, taken from the debate on the Representation of the People Bill of 1948:

“But, as I have said, there was no contract and no bargain made at the Speaker’s Conference to the effect that the present constituencies were going to be perpetuated and that the agreements were going to be perpetuated forever, because the world is moving *quicker today than ever before*, and this Parliament must move with it.”<sup>27</sup>

The formulation of the sequences depends on one’s own research interests and on assumptions about their probable yield in terms of frequency or relevance. These assumptions, however, may be deceptive, e. g., the sequence **{AM} [jɹ\*] \* {AM}** seems to render a high output, but in fact offers only 1021 sentences in 200 years of House of Commons’ debates. This figure appears

---

27 Arthur Woodburn (LAB), Presentation of the People Bill [sic!]. HC Deb 17 February 1948 vol 447 c1021.

high, and there are many highly interesting sentences among the 1021 hits, but in terms of the amount of data available in the HC the figure obviously involves only a very small fraction of all temporal comparisons contained in the corpus. This finding illustrates that several modifications and search queries are needed for each formalized sequence to provide a sufficiently high number of sentences.

Our second formalized sequence uses superlatives and similar terms hinting at the highest possible degree of something. With regard to temporal comparisons the words *never* combined with a time reference (**never** \* **{AM}**) and *ever* combined with a superlative adjective (**{jt}\*** \* **ever**) are the most obvious cases in point. By entering the latter sequence and searching for phrases spoken in the House of Commons displayed as *list*, it is striking to see that superlative phrases of this type are mostly used to highlight presented numerical figures or the high quantitative amount of other entities, whether in a positive or negative way. Here is a typical example sentence taken from a debate on food prices in 1955:

“I am asking if he will state just how far stabilisation and reduction of prices has taken place. Is he not aware of the fact that he has not mentioned that at all? Prices have been going up and up and have reached the *highest level ever* in peace or war.”<sup>28</sup>

Rhetorically speaking, superlative expressions such as these seem not to require any evidence and a particular *comparatum* but are based merely on statements that demand to act on the matter.

Our third formalized sequence contains at its center conjunctions or temporal adverbs that serve as indicators of temporal comparisons. For example, the query as **{j}\*** as **{AM}** may render sentences in which this sequence constitutes only the first part of a more complex comparison, as in the following example, taken from a debate on the international situation in September 1944:

“During this war America, Russia and Britain have made common sacrifices. If we trust each other, as we are doing now, why cannot we trust each other when the war is over, for the future of peace and for the good of the world?”

---

28 Arthur Lewis (LAB), Question on Food Prices. HC Deb 24 November 1955 vol 546 c1638.

Unless we are prepared to do that, and join the common pool, and set up an international police force on the lines I have indicated, then *as sure as night follows day*, there will be another war – perhaps in the next generation.”<sup>29</sup>

That means that our formalized sequences do not always need to contain the complete comparison, but that it may be sufficient to identify parts of it, or hints at their existence, in the sentence before or after. The same observation applies to the sequence **{AM}** \* **not yet** as well, as in the following example, taken from a statement of David Lloyd George in a debate on a possible better future for Russia beyond Bolshevism in April 1919:

“You cannot carry on a great country upon rude and wild principles such as those which are inculcated by the Bolsheviks. When Bolshevism, as we know it and as Russia to her sorrow has known it, disappears, then the time will come for another effort at re-establishing peace in Russia. But that *time is not yet*. We must have patience, and we must have faith. You are dealing with a nation which has been misgoverned for centuries, and been defeated and trampled to the ground, largely, let us admit, owing to the corruption, the inefficiency, and the treachery of its own governors. Its losses have been colossal. All that largely accounts for the real frenzy that seized upon a great people. That is why a nation which has gone through untold horrors has abandoned itself for the moment to fantastic and hysterical experiments. But there are unmistakable signs that Russia is emerging from the trouble. When that time comes, when she is once more sane, calm, and normal, we shall make peace in Russia. Until we can make peace in Russia, it is idle to say that the world is at peace.”<sup>30</sup>

Our fourth formalized sequence takes advantage of the HC’s capacity to search synonyms of certain words. As explained above, in the HC it is no longer necessary to create a list with possible synonyms, but there is a function to search directly for all synonyms. As a matter of course, not every synonym helps to identify comparisons. For example, the sequence **{AM}** \* **[[=change]]**

---

29 John Joseph Tinker (LAB), War and international situation. HC Deb 28 September 1944 vol 403 c549.

30 David Lloyd George (LIB), Question on a Barrier against Bolshevism. HC Deb 16 April 1919 vol 114 c2944-5.

provides a large number of sentences containing the terms *amendment* and *trade* that are in certain contexts used as synonyms of *change*. In these cases there are of course no comparisons. However, this problem is compensated by displaying the found sentences as *list*. The upper right section offers sequences sorted by frequency, and particular expressions can be selected from the list. Further useful expressions of change with time reference are **used to [v\*] {AM}** and **than \* used to be**.

Our fifth formalized sequence tries to identify temporal comparisons in which terms denoting comparisons are used explicitly or which contain particular words that refer to a relation between entities, words such as ‘superior to’. For that type of query it is sufficient to enter just the relevant word with an asterisk in the box *word(s)* and the abbreviation for time reference **{AM}** in the box *collocates*. Depending on whether equations or differentiating comparisons are in the center of consideration, all or part of the following vocabularies may be important for this task: **analog\***, **compar\***, **contrast\***, **simil\***, **inferior\***, **superior\***, **progress\***, etc. Another word that belongs to the semantic field of comparison but needs more elaborate queries is ‘same’. One possibility could be **{AM} \* same \* {AM}**. For a simple exploration of synonyms of ‘comparison’ HC offers the synonym construction **[[=compare]]** to extend one’s own assumptions regarding possible meanings of related words.

Our sixth formalized sequence tries to collect sentences in which comparisons are contested, modified, or otherwise made a topic of debate. Expressions that may indicate a contestation of comparison or lead to new comparisons substituting old ones are, for example, \* **[be] \* [[difference] | [discrepance] | [change] | [distinction] | [dissimilarity]]** or \* **[be] \* analogy**. Here is an example sentence, taken from a debate on the Proportional Representation Bill of 1921:

“At the present time the electors of all parties are glad to put confidence in and to seek help and advice from their Member, because they regard him as ‘Our Member.’ They know who he is. Under the new system it will be absolutely impossible for a candidate at an election to visit every polling district in a huge area, and it will be very difficult, indeed, for him to remain in touch with the whole of the wide and scattered constituency after he has become elected. *There is no analogy or comparison* at all between the case of a Member

of Parliament and that of a guardian elected for a little town in Ireland. I am not a bit impressed by those arguments from Ireland."<sup>31</sup>

The example shows a rejection of comparability, and in this case a rejection that uses a temporal comparison to justify the statement. However, without a close reading of the sentence in context it would have been impossible to identify the temporal components of the comparison. The time references *at the present time* and *new system* are uttered two sentences before the searched sequence and thus beyond the typical collocation span. This makes it nearly impossible to formalize such sequences in advance.

The application of all six formalized sequences works quite well and provides a large amount of comparison-performing speech acts. In later steps of our project the formalized sequences may help to identify comparisons in other text corpora, particularly in those that are dominated by argumentation and persuasion. For a detailed historical semantic analysis the number of sentences resulting from our searches in the HC is far too big to analyze them historically. At this point it is essential to apply limits to the inquiry. Different options are available in the HC to achieve this. One possibility entails bringing the functioning of HC closer to traditional historical research by narrowing the sample by time, speaker, party and other settings. For questions focused on actors, the HC offers not only a limitation by party membership of the speaker but also a party membership related to the governing party. This search function is helpful for examining shifting usages of concepts or patterns of argument within a party over time, between political parties, or between the roles of government and opposition. Another possibility to build a more specific sample is to create a list with relevant speakers and to use the corpus of all their speeches or to select particular speeches of different speakers. For that purpose, the section *decade/speaker* contains the option 'create list' where one can set several parameters – chamber, speaker, time period and party.

As already mentioned, the HC does not allow to select debates by subject matter in order to find out, for example, in which thematic contexts comparisons are most likely to appear. However, our sixth sequence that relates to comparisons that are contested or modified or otherwise made a topic may

---

31 Gerald Hurst (CON), Proportional Representation Bill. HC Deb 05 April 1921 vol 140 c637.

help to identify these contexts. In addition, our fifth category, referring to the terms denoting comparison (compar\*, etc.) and their collocations, may also be useful for this purpose.

### 3.3 Between distant and close reading

From a historian's perspective, there are advantages and disadvantages of working with the HC. The most notable advantage is of course the sheer amount of data available in the HC and the web server's capacity to allow fast searches.<sup>32</sup> Searches for single words or collocations that may be historically relevant in political contexts can be effectuated on a large scale and with unprecedented speed. For conceptual historians who want to test hypotheses about occurrences of particular words or combinations of words, this is an invaluable tool. Beyond such simple searches, the time required for 'learning' to formulate precise queries and for overcoming disappointments in several sessions of 'trial and error' has to be measured against the possible gains in terms of surprise findings that would be impossible for anyone reading the debates in a traditional, hermeneutical way. Working with the HC and its inbuilt tagger (HTST) thus requires a willingness to switch frequently between the techniques of distant reading and subsequent close reading, or the other way round. Moreover, the lack of a search function by subject matter of debates requires various detours in order to create a more limited text corpus that may be better suited to respond to the historian's particular interests. Two other – very practical – disadvantages are the limitation of access and various restrictions for saving the relevant text materials.

Despite the disadvantages, however, the uniqueness of the HC lies not only in the collection of almost all parliamentary speeches from 200 years but in the multiple annotations of every single lexical unit in the corpus. Therefore, the HC optimally supports 'traditional' conceptual-history searches of single terms and their collocations, and, due to the precise documentation of speakers, parties, political roles, dates and debates, it also provides important information to contextualize the findings; this is particularly helpful for researchers interested in the conceptual horizons of historical actors. Fur-

---

32 Anthony, Laurence, A critical look at software tools in corpus linguistics, in: *Linguistic Research* 30 (2013), 152–153.



thermore, the HTST allows very elaborated queries in the sense of parts of speech, patterns of sentences, rhetorical figures, punctuation and further criteria. Even if the exceptionality of British parliamentary language has to be kept in mind, the Hansard debates are thus an outstanding source for the history of political language and rhetoric in general.

#### 4. Case study II: man vs. machine – towards a universal tool for identifying comparisons

The literary genre of utopias should be a good source for identifying comparisons, as it inherently deals with differences. Already in Thomas More's eponymous *Utopia* (1516),<sup>33</sup> which deals with the differences between European nations and the imagined state of the island Utopia, comparisons have the function of closing the gaps between the fictional storyline and a potential reader who is situated in his or her own present. In the subgenre of uchronia, which owes its name to Charles Renouvier's utopian novel *Uchronie* (1876),<sup>34</sup> the problem of closing gaps in space and time via comparison is even more complex. If the genre as a whole is apparently depending on the use of comparisons in order to function as a literary text, it follows that comparisons, especially temporal comparisons, should appear quite frequently in utopias and uchronias.

The importance of utopias and uchronias should be evident for historians. In general, time is the dimension historians are used to work with; according to Niklas Luhmann this is due to the function of synchronization, which historiography fulfills in a differentiated social system.<sup>35</sup> And according to Luhmann, as well as Reinhart Koselleck, time is multidimensional, consisting of pasts, presents and futures, which are interwoven.<sup>36</sup> Thus, in the last forty years, the 'futures past' have been a research topic for historians in the

---

33 More, Thomas, *Utopia*, trans. Gilbert Burnet [1551], New York: Cassel & Co., 1901.

34 Renouvier, Charles, *Uchronie – L'Utopie dans l'Histoire*, Paris: Bureau de la critique philosophique, 1876.

35 See Luhmann, Niklas, *Weltzeit und Systemgeschichte: Über Beziehungen zwischen Zeithorizonten und sozialen Strukturen gesellschaftlicher Systeme*, in: *Soziologie und Sozialgeschichte, Special Issue* 16 (1973), 81–115.

36 Koselleck, Reinhart, *Zeitschichten: Studien zur Historik*, Frankfurt a. M.: Suhrkamp, 2000.

same way as the bygone past, and in that context the literary genres of utopia and uchronia have become a privileged source for historians to access the interplay of time-layers which are so visible in them.

For the purposes of this article we have put together a small text corpus of utopias/uchronias that had to fulfill two basic conditions. First, for pragmatic reasons and better comparability of results, all texts had to be in English, whether in the original or a (contemporary) translation. Secondly, the corpus had to cover a long time span from the sixteenth century to the present. Obviously the first text to consider was Thomas More's *Utopia*. In addition, a small selection of the classics of the genre had to be considered as well, starting with the unfinished manuscript of Francis Bacon's *New Atlantis* (1627) and James Harrington's *The Commonwealth of Oceana* (1656), which are both, like More's *Utopia*, discourses of state philosophy.<sup>37</sup> Also added to the corpus was the English translation of Louis-Sebastien Mercier's uchronia *The Year 2500* (1795, the French original *L'an 2440* appeared in 1771).<sup>38</sup> Reinhart Koselleck considered Mercier's *L'an 2440* to be the very first uchronia, but this assumption can no longer be maintained since there are at least three earlier texts by Samuel Madden (1733), Heinrich Gottlob Justi (1759) and an anonymous author (1763) which may well be described as uchronias (but these rather obscure texts were not included in the corpus).<sup>39</sup> The next example to be included was the first dystopia, the inversion of the positive state or possible future-to-be to the negative: Mary Shelley's *The Last Man* (1825).<sup>40</sup>

---

37 Bacon, Francis, *The advancement of learning and New Atlantis* [1627], London: Oxford University Press, 1969.

38 Mercier, Louis-Sebastien, *Memoirs of the Year 2500*, trans. Thomas Dobson [1795], Boston: Gregg Press, 1977.

39 See: R. Koselleck, *Zeitschichten*, 131; Madden, Samuel, *Memoirs of the Twentieth Century*, London: Unknown, 1733; Justi, Heinrich Gottlob, *Untersuchung ob etwan die heutigen europäischen Völker Lust haben möchten, dereinst Menschenfresser oder wenigstens Hottentotten zu werden*, Philadelphia: Jacob Heinrich Lowe, 1759; Anonymus, *The Reign of Georg VI: 1900–1925*, London: Printed for W. Niccoll at the Paper Mill in St. Paul's Churchyard, 1763, ed. C. Oman (London: Reprinted by Rivingtons, 34, King Street Covent Garden, 1899).

40 Shelley, Mary, *The Last Man* (1825), ed. Hugh Luke, Lincoln: University of Nevada Press, 1965.

Following in chronological order, there are three thematic clusters. The first cluster is formed by novels of a socialistic future, which were popular at the end of the nineteenth century on both sides of the Atlantic. This cluster contains Edward Bellamy's *Looking Backward: 2000–1887* (1888) and Jack London's *The Iron Heel* (1908), which is written like a dystopia, but is actually a utopia. Also part of that cluster are William Morris's *News from Nowhere* (1890) and Eugen Richter's libertarian answer *Pictures of a Socialistic Future* (1893, the German original *Sozialdemokratische Zukunftsbilder* appeared in 1890).<sup>41</sup> Aldous Huxley's *Brave New World* (1932) is included as a stand-alone, because it is neither certain whether this is a utopian or dystopian novel nor does it fit into any of the thematically organized clusters.<sup>42</sup>

The second cluster, which deals with the rise of communism and fascism, was created around George Orwell's *1984* (1949). It also contains Yevgeny Zamyatin's *WE* (1924), which Orwell used as a template for *1984*, Karin Boye's *Kalocain* (Swedish original 1940, English translation 1966) and the Post-WWII *Fahrenheit 451* (first published 1953) by Ray Bradbury.<sup>43</sup>

With Murray Leinster's short story *A Logic Named Joe* (1946) begins the third thematic cluster, which deals with the potentially negative influence of computer technology on human society. The same applies to Isaac Asimov's *I, Robot* (1950), Philip K. Dick's *Do Androids Dream of Electric Sheep?* (1968), William Gibson's *Neuromancer* (1984), and Adam Sternbergh's *Shovel Ready* (2014).<sup>44</sup> An exception of a kind is Ernest Callenbach's *Ecotopia* (1975), which is the only yet found utopian novel of the second half of the 20th

---

41 Bellamy, Edward, *Looking Backward: From 2000 to 1887, 1888*, Cambridge, Massachusetts: Harvard University Press, 1978. London, Jack, *The Iron Heel*, 1908, Auckland: Floating Press 2009. Morris, William, *News from Nowhere, or an Epoch of Rest, being some chapters from A Utopian Romance*, 1890, New York, Bombay and Calcutta: Longmans, Green & Co. 1908. Richter, Eugene, *Pictures of a Socialistic Future*: Freely adapted from Bebel, London: Swan Sonnenschein, 1893.

42 Huxley, Aldous, *Brave New World* [1932], Middlesex: Penguin Books, 1971.

43 Orwell, George, 1984, London: Secker & Warburg 1949; London: Penguin, 2016. Zamyatin, Eugene, *WE*, trans. Gregory Zilboorg, New York: E. P. Dutton, 1924; New York: E. P. Dutton 1952. Boye, Karin, *Kalocain*, trans. Gustav Lannestock, Madison, Wisconsin: University of Wisconsin Press, 1966. Bradbury, Ray, *Fahrenheit 451*, Philadelphia: Chelsea House Publ., 2001.

44 Leister, Murray, *A Logic Named Joe*, in: *Astounding Science Fiction* 37 (1946), 139–154. Dick, Philipp K., *Do Androids Dream of Electric Sheep?*, New York: Ballantine, 1968. Asimov, Isaac, *I, Robot*, New York, Doubleday & Co., 1950. Gibson, William, *Neuromancer*,

century that deals with practical solutions to the upcoming ecological crisis.<sup>45</sup> Anthony Burgess's *A Clockwork Orange* (1962) is also part of the corpus, but will be ignored in this study, due to the used sociolect with its own dictionary.<sup>46</sup>

Once the corpus was put together, there were two possible options to search for (temporal) comparisons. The first option was to simply read the novels, mark all the sentences that include comparisons 'by hand', and transcribe them into a document. The second option was to digitize all texts and apply one or more search tools to identify (temporal) comparisons in the corpus. Since the digitization and adaptation of available tools in collaboration with the INF project proved to be time-consuming, both options were pursued in parallel; therefore this case study has amounted in fact to an experimental comparison between the outcome and relative efficiency of the work of 'man and machine'.

Reading those novels was certainly not 'simple' reading, but it implied the careful questioning and categorizing of every sentence in order to decide whether it actually was (or contained) a comparison or not. This kind of focused reading took a lot of time. The 2151 pages<sup>47</sup> of the corpus took approximately 143 hours of focused reading, by an average reading speed of 15 pages per hour. It was comforting to know, however, that there would be no way around reading the novels anyway, because the analysis of the functions and uses of comparisons in the literary texts cannot be done without careful scrutiny. Still, the process of marking the sentences and transferring them into digital data was a task which a machine could certainly have done much faster than a human. My estimate is that this process, if done by a machine, would have saved more than 60 % of the invested time because in that case the reading speed could have been increased to 40 pages per hour.

---

New York: Ace Books, 1984. *Sternbergh, Adam*, Shovel Ready, New York: Random House, 2014.

45 *Callenbach, Ernest*, *Ecotopia: The Notebooks and Reports of William Weston*, Berkeley: Banyan Tree Books 1975.

46 *Burgess, Anthony*, *A Clockwork Orange: The Restored Edition*, London: Random House, 1962; London, Random House, 2012.

47 Based on the printed DinA4-Pages of the .txt-documents.

## 4.1 Translating for a machine

Before a machine can be employed there is the problem of text format. Most of the novels which were written before 1950 were available on archive.org or Gutenberg.org, normally already in .txt-format. In that case the texts only had to be ‘cleansed’ of headers, rights-disclaimers, introductions of the editors, etc. The other texts had to be scanned to PDF first, then converted into .txt-format, and from there into .xml-format to then be tagged. The tagger used for all documents was USAS, an open source tagger that does lemmatization and semantic annotations. The tagging procedure is necessary to enable the ‘machine’ to search for more than just individual signs, e. g., for inflected adjectives or for words semantically tagged as time references. As a query-processing-software we used the *IMS Corpus Workbench* in the web-based version (CWB).<sup>48</sup> This query-processing software is a search-engine that allows inserting simple or complex syntax-based patterns and shows the results on the screen.

Working with such a tool demands to formalize syntactic sequences beforehand, in our case sequences that express temporal comparisons. At this stage, it proved to be tremendously helpful that the entire corpus had already been read and the temporal comparisons marked ‘by hand’. Thus, several hundred temporal comparisons had already been identified and classified.

Among these, two main types were particularly salient. The first type may be called inter-timeline-comparison: One or more *comparata* X situated in a time A ( $T=0$  or  $T=0 \pm x$ ) are compared to one or more *comparata* Y situated in a time B ( $T=0$  or  $T=0 \pm x$ ) with regard to one or more *tertia* V. An example sentence for type I, taken from Ernest Callenbach’s *Ecotopia*, would be:

“In Ecotopia (time A) at that time (time A; indicator  $T=T-x$ ), as in the United States (time B) now (time B; indicator  $T=0$ ), such areas (comparata) were

---

48 The Claws7-tagset and USAS-semantic-tagset were developed at the University Centre for Computer Corpus Research on Language (UCREL) at Lancaster University. The open source query-processing software was also developed at UCREL. For more information see [ucrel.lancs.ac.uk](http://ucrel.lancs.ac.uk) [accessed: 01.09.2019].

mainly devoted to (tertium=areal usage) factories, warehouses, sewage plants, railroad yards, dumps, and other unsavory uses.”<sup>49</sup>

The second type may be called inner-timeline-comparison. In the utopian genre this type of comparison is often used when characters in the novel reflect upon themselves and their development across time: One *comparatum* X situated in a present (T=0) is compared to the same *comparatum* X at another point on a timeline (T=T +/- x) with regard to one or more *tertia* V. An example sentence for type II, taken from Eugen Richter's *Pictures of a Socialistic Future* would be:

“I (comparatum) am sorry to say that I can now (T=0) no longer (indicator T=T-x) take my meals (tertium=meal times) with my wife except on Sundays, as I have been accustomed for the last (indicator T=T-x) twenty five years (x=25 years).”<sup>50</sup>

Defining the *comparata* in a search list beforehand is hardly possible; they may be characters in the novel, groups, places, objects, abstract thoughts or historical periods. However, our traditional reading and marking procedure has shown that comparative references to the ancient Greeks or Romans, to “the savage”, “caveman” or “primitive man”, to earlier stages of civilization, to an “other world” or an “old(er) world”, to something “medieval” or “ancient”, appear quite frequently in utopias. Similarly, in inner-timeline comparisons, “boyhood”, “childhood”, “boy” and “girl” are often used as comparative references for the novel character at his or her respective present (T=0).

Much easier to define beforehand are the temporal indicators. They comprise temporal prepositions or adverbs like before, last (day, month, year, century, era, seasons etc.), yore, formerly, heretofore, (than) ever, since, earlier, etc., but also possessive pronouns followed by the word “time” (In my time ..., In your time.), as well as specified or unspecified temporal relations (earlier time/that time) and the expression “used to be”. The indicators for the respective present are relatively few, mostly T=0 is represented by the word “now”, followed by “no longer” or “no more”. Sometimes, however, a reference

49 E. Callenbach, *Ecotopia*, 87.

50 E. Richter, *Pictures of a Socialistic Future*, 43.

to a present may be hidden in a personal pronoun like “we”, as in a sentence such as “*In the 19th century... We...*” or even by simply opposing “*You ... We...*” without any temporal indicator at all. While someone who reads in the traditional, ‘analogue’ style will often be able to detect the temporal comparison that is hidden in such an opposition, it seems unlikely that a machine can ever be trained to such a degree that it will correctly select these and similarly ‘hidden’ temporal comparisons.

Having established two main types of temporal comparison and described their way of functioning in utopian novels, we may now proceed to formulate queries in the language used by Corpus Workbench. In the following table the first column shows the syntactical query, while the second column explains its respective aims. All queries were constructed with the general or semantic tagset, and they include most of the above mentioned indicators. Not included are references to different stages of civilization, so misses are to be expected.

*Table 2: Set of queries with taggers*

QNo	Query	Description
Q1	[sem contains "T1.1.1"][*] [sem contains "T1.1.2"] within s;	Searches for a semantically tagged word relating to the past and a semantically tagged word relating to the present within a sentence.
Q2	[sem contains "T1.1.1"][*] [sem contains "T1.1.2"] within 50;	Searches for a semantically tagged word relating to the past and a semantically tagged word relating to the present within a range of 50 words.
Q3	[sem contains "T1.1.1"][*] [sem contains "T1.1.2"] within 75;	Searches for a semantically tagged word relating to the past and a semantically tagged word relating to the present within a range of 75 words.
Q4	[sem contains "T1.1.2"][*] [pos="RRR"][*] [sem contains "T1.1.1"] within s;	Searches for a semantically tagged word relating to the past, followed by comparative general adverb and a semantically tagged word relating to the present within a sentence.
Q5	[sem contains "T1.1.2"][*] [pos="RRR"][*] [sem contains "T1.1.1"] within 50;	Searches for a semantically tagged word relating to the past, followed by comparative general adverb and a semantically tagged word relating to the present within a range of 50 words.

QNo	Query	Description
Q6	[sem contains "T1.1.2"][*] [pos="RRR"][*][sem contains "T1.1.1"] within 75;	Searches for a semantically tagged word relating to the past, followed by comparative general adverb and a semantically tagged word relating to the present within a range of 75 words.
Q7	[sem contains "T1.1"][*][sem contains "A6"] within s;	Searches for a semantically tagged word relating to the past and a general comparative term within a sentence.
Q8	[sem contains "T1.1"][*][sem contains "A6"] within 50;	Searches for a semantically tagged word relating to the past and a general comparative term within a range of 50 words.
Q9	[sem contains "T1.1"][*][sem contains "A6"] within 75;	Searches for a semantically tagged word relating to the past and a general comparative term within a range of 75 words.
Q10	[pos="RRR"][*][pos="NNT1"]  [pos="NNT2"] within s;	Searches for a comparative adverb, followed by a temporal noun, in singular or plural, within a sentence.
Q11	[pos="RRR"][*][pos="NNT1"]  [pos="NNT2"] within 50;	Searches for a comparative adverb, followed by a temporal noun, in singular or plural, within a range of 50 words.
Q12	[pos="RRR"][*][pos="NNT1"]  [pos="NNT2"] within 75;	Searches for a comparative adverb, followed by a temporal noun, in singular or plural, within a range of 75 words.

## 4.2 Counting for a machine

Once the search queries had been formulated they could be put to a test. To make things a little bit easier, we selected only five utopian novels: Thomas More's *Utopia* as the oldest, Adam Sternbergh's *Shovel Ready* as the most recent one, and in between one for each of the thematic clusters: Jack London's *The Iron Heel* representing the socialism-cluster, George Orwell's *1984* representing the fascism/communism-cluster and Philip K. Dick's *Do Androids Dream of Electric Sheep?* for the fear-of-the-machine-cluster. For a first attempt, this selection seemed sufficiently large to represent different variants of English and of temporal comparison.

The first test served to determine the success rate of the machine in identifying those comparisons that had already been found 'by hand'. The



following table shows the total number of temporal comparisons identified ‘by hand’ in the column next to the book title, followed by the result for each query, the total number for all queries, and in the last column the success rate for the machine in percent.

Table 3: Comparison of results man vs. machine

Title	Total by hand	Q1	Q2	Q3	Q4	Q5	Q7	Q8	Q9	Q10	Q11	Q12	Total machine	%
Utopia	19	0	0	0	0	0	0	0	0	0	0	0	0	0
Iron Heel	71	1	0	5	0	0	0	0	0	8	0	0	14	19.72
1984	50	2	0	12	1	0	0	0	0	9	0	0	22	44
Ecotopia	35	2	3	0	0	0	0	0	0	0	0	0	5	14.29
Shovel Ready	39	0	0	3	0	0	0	0	0	0	0	0	3	7.69
Total	214	5	3	20	1	0	0	0	0	17	0	0	44	20.56

The results are extremely disappointing for our digital tools. They only found 44 temporal comparisons out of the 214 found ‘by hand’. This amounts to a very low overall success rate of 20.56 percent, which seems to indicate that the software is not worth giving it further tries. It worked still reasonably well for *1984* but had very poor results for *Utopia* and *Shovel Ready*. Why there were no hits at all for *Utopia* is hard to understand. It may either be that there was a problem in the pipeline, i. e., in the process of formatting and tagging, or that the maximum search range of 75 words is too small because many of the temporal comparisons are made within a longer sentence or across several long sentences. The main reason why there were so few hits in *Shovel Ready* could be the high context dependence of temporal comparisons within the story and the short and often incomplete sentences that are used. The story of *Shovel Ready* builds on a terror act with a dirty bomb which alters everyday life in New York City in the near future. Spademan, the protagonist, a former garbageman and now a contract killer, lost his wife in the attack. In the limnopsphere, a more progressive form of the internet, he encounters a digital version of his wife again. “My wife. In the same dress I last saw her in.” The reader knows that his wife is dead and that the scene takes place after the attack. The machine should have recognized this as well because the semantic

tagger T.1.1.1 should have identified ‘last’ as a word relating to the past and ‘same’ should have been identified by the tagset category A6 as part of a comparison. So the fault does not seem to be a result of the novel’s specific language, but rather indicates a problem with the tagset, which does not work as accurately as hoped for. But one should stress that it did not work as badly as the table indicates. For the above table does *not* include those comparisons that the machine found *in addition* to those found by manual search. These are now shown in table 4:

*Table 4: Additional temporal comparisons found by the machine*

Title	Total
Utopia	0
Iron Heel	15
1984	29
Ecotopia	13
Shovel Ready	1
Total	58

These results show that neither man nor machine are perfect. There are misses on both sides. Apparently, the machine is unable to ‘understand’ the story which is created by the text as a whole and will therefore miss temporal comparisons that are obvious for human readers. On the other hand, it may help us to compensate for gaps in concentration during the process of reading. The tagsets which come along with Corpus Workbench are far from perfect.

Better results might be achieved in the future by coding one’s own semantic tagsets. One could do this either by editing the existing Corpus Workbench tagset. Or – and that is another test we have made – by creating searches that use our own above mentioned indicators. Admittedly, this is a much less elegant solution, but it turned out that it worked a little bit better than the tagset. On such a basis an archetype for a different kind of semantic tagger could be created. Table 5 shows the set of queries used for this test.

Table 5: Set of queries with our own indicators

QNo	Query	Description
Q13	"before"   "past"   "last"   "century"   "yore"   "formerly"   "heretofore"   ("than" [* "ever"])   "caveman"   "savage"   "era"   "your"   "time"   "since"   "earlier"   "no"   "more"   "no"   "longer"   "primitive"   "man"   "now"   "ancient"   "roman"   "greek"   "to-day"   "today"   "old"   "years"   "month"   "day"   "never"   "modern"   "history"   "first"   "once"   "present" within s;	Searches for a set of indicators within a sentence.
Q14	"before"   "past"   "last"   "century"   "yore"   "formerly"   "heretofore"   ("than" [* "ever"])   "caveman"   "savage"   "era"   "your"   "time"   "since"   "earlier"   "no"   "more"   "no"   "longer"   "primitive"   "man"   "now"   "ancient"   "roman"   "greek"   "to-day"   "today"   "old"   "years"   "month"   "day"   "never"   "modern"   "history"   "first"   "once"   "present" within 50;	Searches for a set of indicators within a range of 50 words.
Q15	"before"   "past"   "last"   "century"   "yore"   "formerly"   "heretofore"   ("than" [* "ever"])   "caveman"   "savage"   "era"   "your"   "time"   "since"   "earlier"   "no"   "more"   "no"   "longer"   "primitive"   "man"   "now"   "ancient"   "roman"   "greek"   "to-day"   "today"   "old"   "years"   "month"   "day"   "never"   "modern"   "history"   "first"   "once"   "present" within 75;	Searches for a set of indicators within a range of 75 words.

Again, these queries were applied to the same five utopias mentioned above. This time, the column next to the book title contains the total of all comparisons found manually and those that the machine found in addition. The results are as follows.

Table 6: Comparing results of man vs. machine with our own indicators

Title	Total	Q13	Q14	Q15	Total	%
Utopia	19	0	0	0	0	0
Iron Heel	86	64	2	0	66	76.74
1984	79	32	4	0	36	45.57
Ecotopia	38	26	0	0	26	68,2
Shovel Ready	40	6	1	0	7	17.0
Total	262	128	7	0	135	51.53

Searching only for our self-defined indicators of temporal comparisons led to much better results than the queries based on the tagsets. With the exception of *Utopia*, which still did not show any valid result, the overall success rate of 51.53 percent is, as expected, much higher than that produced by the application of the tagsets. One might argue that this method is a kind of self-delusion considering that we are searching for something that is already known to be there. The next step might be to apply our self-defined search queries to a different text corpus and compare the results with those achieved by a traditional reading taking place after that experiment.

### 4.3 And the winner is ...

Overall there is much work left to do. At this stage of our cooperative project, we are still very far away from our desired aim of disposing of a 'machine' that can reliably identify temporal comparisons in the English language. We will either have to modify the Corpus Workbench's tagsets or construct our own tagger for indicators of temporal comparisons. Using a syntax which only relies on the indicators would be a third and not so elegant way to solve this problem, but even this would still require a revision of the indicators and a test with another corpus.

## 5. Conclusion

Our case studies on the already digitized and tagged Hansard Corpus on the one hand and on the self-defined corpus of utopias on the other hand have shown that the successful identification of temporal comparisons in texts by means of digital tools is highly dependent on the quality of the tagsets used. Whereas the search facilities provided by the HC interface, supported by the HTST tagger, have rendered some valuable results, the rate of 'hits' produced by a freely available tagger (Corpus Workbench) applied to the corpus of utopias was rather low, especially if put into relation to the amount of time needed to prepare the textual material and to adapt the codes used by the tagset to our specific questions. Even if the problems discussed in this article can be solved in a satisfactory way, it may still be a long way to go from the reliable identification of certain types of comparison-performing utterances to a historical analysis of their changing use patterns in certain historical

contexts or literary genres. However, there is no way of denying the progress in terms of acceleration brought by searchable corpora and elaborated taggers like the HTST, compared to the efforts and time that would have been necessary in the pre-digital age to effectuate a traditional conceptual history of 'comparison', its adjacent concepts, and basic forms of articulating comparisons. The result of our explorative inquiry, therefore, is in itself a temporal comparison, and one that is advantageous for the present compared to a not too remote past.

## Bibliography

- Alexander, Marc/Davies, Mark*, Hansard Corpus 1803–2005, [https://www.hansard-corpus.org/help/syntax\\_e.asp](https://www.hansard-corpus.org/help/syntax_e.asp) [accessed: 29.04.2019].
- Anonymus*, The Reign of Georg VI: 1900–1925, London: Printed for W. Niccoll at the Paper Mill in St. Paul's Churchyard, 1763.
- Anthony, Laurence*, A critical look at software tools in corpus linguistics, in: *Linguistic Research* 30 (2013), 152–153.
- Asimov, Isaac*, I, Robot, New York, Doubleday & Co., 1950.
- Bacon, Francis*, The advancement of learning and New Atlantis [1627], London: Oxford University Press, 1969.
- Bellamy, Edward*, Looking Backward: From 2000 to 1887, 1888, Cambridge, Massachusetts: Harvard University Press, 1978.
- Bossuet, Jacques-Bénigne*, Discours sur l'Histoire universelle [1681], Paris: Garnier-Flammarion, 1966.
- Boye, Karin*, Kallocain, trans. Gustav Lannestock, Madison, Wisconsin: University of Wisconsin Press, 1966.
- Bradburry, Ray*, Fahrenheit 451, Philadelphia: Chelsea House Publ., 2001.
- Bunyan, John*, The Jerusalem Sinner Saved [1688], Edinburgh: The Banner of Truth, 2005.
- Burgess, Anthony*, A Clockwork Orange: The Restored Edition, London: Random House, 1962; London, Random House, 2012.
- Callenbach, Ernest*, Ecotopia: The Notebooks and Reports of William Weston, Berkeley: Banyan Tree Books 1975.
- Carroll, Lewis*, Alice's Adventures in Wonderland [1865], Los Angeles: Enhanced Media, 2016.

- Cheah, Pheng*, The Material World of Comparison, in: *New Literary History* 40 (2009).
- Descartes, Rene*, *Ceuvres*. T. X, Paris: Cerf, 1908.
- Dick, Philipp K.*, *Do Androids Dream of Electric Sheep?*, New York: Ballantine, 1968.
- Eggers, Michael*, *Vergleichendes Erkennen: Zur Wissenschaftsgeschichte und Epistemologie des Vergleichs und zur Genealogie der Komparatistik*, Heidelberg: Universitätsverlag Winter, 2016.
- Gibson, William*, *Neuromancer*, New York: Ace Books, 1984.
- Grafton, Anthony*, Comparisons Compared: A Study in the Early Modern Roots of Cultural History, in: Renaud Gagné/Simon Goldhill/Geoffrey E. R. Lloyd (eds.), *Regimes of Comparatism: Frameworks of Comparison in History, Religion and Anthropology*, Leiden/Boston: Brill, 2019, 18–48.
- Grave, Johannes*, Vergleichen als Praxis. Vorüberlegungen zu einer praxis-theoretisch orientierten Untersuchung von Vergleichen, in: Angelika Epple/Walter Erhart (eds.), *Die Welt beobachten: Praktiken des Vergleichens*, Frankfurt a. M.: Campus, 2015, 136–137.
- Huxley, Aldous*, *Brave New World* [1932], Middlesex: Penguin Books, 1971.
- Jakobson, Roman*, The Cardinal Dichotomy of Language, in: Ruth Nanda Anshen (ed.), *Language: An Enquiry into its Meaning and Function*, Port Washington and London: Kennikat Press, 1971.
- James, William*, *Pragmatism and Four Essays from 'The Meaning of Truth'*, New York: Meridian Books, 1955.
- Justi, Heinrich Gottlob*, *Untersuchung ob etwan die heutigen europäischen Völker Lust haben möchten, dereinst Menschenfresser oder wenigstens Hottentotten zu werden*, Philadelphia: Jacob Heinrich Lowe, 1759.
- Kay, Christian*, Diachronic and synchronic thesauruses, in: Philip Durkin (ed.), *The Oxford Handbook of Lexicography*, Oxford: Oxford University Press, 2015, 367–380.
- Koselleck, Reinhart*, *Zeitschichten: Studien zur Historik*, Frankfurt a. M.: Suhrkamp, 2000.
- Leister, Murray*, A Logic Named Joe, in: *Astounding Science Fiction* 37 (1946), 139–154.
- London, Jack*, *The Iron Heel*, 1908, Auckland: Floating Press 2009.
- Luhmann, Niklas*, *Liebe als Passion. Zur Kodierung der Intimität*, Frankfurt a. M.: Suhrkamp, 1982.

- Luhmann, Niklas*, Weltzeit und Systemgeschichte: Über Beziehungen zwischen Zeithorizonten und sozialen Strukturen gesellschaftlicher Systeme, in: *Soziologie und Sozialgeschichte, Special Issue* 16 (1973), 81–115.
- Madden, Samuel*, *Memoirs of the Twentieth Century*, London: Unknown, 1733.
- Mauz, Andreas/Sass, Hartmut von (eds.)*, *Hermeneutik des Vergleichs. Strukturen, Anwendungen und Grenzen komparativer Verfahren*, Würzburg: Königshausen & Neumann, 2011.
- Mercier, Louis-Sebastien*, *Memoirs of the Year 2500*, trans. Thomas Dobson [1795], Boston: Gregg Press, 1977.
- Mignolo, Walter D.*, Who Is Comparing What and Why, in: Rita Felski/Susan Stanford Friedman (eds.), *Comparison: Theories, Approaches, Uses*, Baltimore: John Hopkins University Press, 2013.
- More, Thomas*, *Utopia*, trans. Gilbert Burnet [1551], New York: Cassel & Co., 1901.
- Morris, William*, *News from Nowhere, or an Epoch of Rest, being some chapters from A Utopian Romance*, 1890, New York, Bombay and Calcutta: Longmans, Green & Co. 1908.
- Orwell, George*, 1984, London: Secker & Warburg 1949; London: Penguin, 2016.
- Palonen, Kari*, Concepts and Debates. Rhetorical Perspectives on Conceptual Change, in: Willibald Steinmetz/Michael Freedden/Javier Fernández Sebastián, *Conceptual History in the European Space*, New York and London: Berghahn, 2017, 96–117.
- Perelman, Chaim/Olbrechts-Tyteca, Lucie*, *The New Rhetoric: A Treatise on Argumentation*, trans. John Wilkinson and Purcell Weaver, 2nd ed., Notre Dame and London: University of Notre Dame Press, 1958.
- Perrault, Charles*, *Parallèle des anciens et des modernes en ce qui regarde les arts et les sciences* [1687], München: Eidos, 1964.
- Piao, Scott et al.*, A time-sensitive historical thesaurus-based semantic tagger for deep semantic annotation, in: *Computer Speech & Language* 46 (2017), 113–135.
- Postoutenko, Kirill*, Preliminary typology of comparative utterances: a tree and some binaries (in print).
- Renouvier, Charles*, *Uchronie – L’Utopie dans l’Histoire*, Paris: Bureau de la critique philosophique, 1876.
- Richter, Eugene*, *Pictures of a Socialistic Future: Freely adapted from Bebel*, London: Swan Sonnenschein, 1893.

- Richter, Melvin*, "That vast Tribe of Ideas". Competing Concepts and Practices of Comparison in the Political and Social Thought of Eighteenth-Century Europe, in: *Archiv für Begriffsgeschichte* 44 (2002), 199–219.
- Shelley, Mary*, *The Last Man* (1825), ed. Hugh Luke, Lincoln: University of Nevada Press, 1965.
- Shirer, William*, *Berlin Diary – The Journal of a Foreign Correspondent, 1934–1941*, New York: Knopf, 1942.
- Shakespeare, William*, *Much Ado about Nothing* [1623], Philadelphia, PA: J. P. Lippincott Company, 2001.
- Sternbergh, Adam*, *Shovel Ready*, New York: Random House, 2014.
- Steinmetz, Willibald*, Introduction: Concepts and Practices of Comparison in Modern History, in: Willibald Steinmetz (ed.), *The Force of Comparison: A New Perspective on Modern European History and the Contemporary World*, Oxford/New York: Berghahn Books, 2019, 1–32.
- Steinmetz, Willibald*, Above/below, better/worse, or simply different? Metamorphoses of Social Comparison, 1600–1900, in: Willibald Steinmetz (ed.), *The Force of Comparison: A New Perspective on Modern European History and the Contemporary World*, New York and Oxford: Berghahn, 2019, 80–112.
- Steinmetz, Willibald*, "Vergleich" – eine begriffsgeschichtliche Skizze, in Angelika Epple/Walter Erhart (eds.), *Die Welt beobachten: Praktiken des Vergleichens*, Frankfurt/New York: Campus Verlag, 2015, 85–134.
- University of Glasgow*, SAMUELS Project (Semantic Annotation and Mark-Up for Enhancing Lexical Searches), <https://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/samuels/> [13.03.2019].
- Voltaire [François-Marie Arouet]*, *Des Juifs* [1756], in: André Versaille (ed.), *Dictionnaire de la pensée de Voltaire par lui-même*, Bruxelles: Editions Complexe, 1994.
- Zamiatin, Eugene*, WE, trans. Gregory Zilboorg, New York: E. P. Dutton, 1924; New York: E. P. Dutton 1952.





## Authors

---

**Anna Dönecke** is a historian of the early modern period specializing in intercultural legal history and works as a PhD within the SFB 1288 “Practices of comparing” at Bielefeld University.

**Michael Götzelmann** specializes in the history of practices of comparing in the genre of utopias from the Middle Ages to today with a focus on Conceptual History working within the SFB 1288 “Practices of Comparing” at Bielefeld University.

**Marcus Hartner** holds a PhD in English Studies and specializes in narratology, early modern travel literature, and intercultural relations between England and the Islamic world. He is currently working at Bielefeld University.

**Joris Corin Heyder** holds a PhD in art history specializing in medieval and early modern art as well as connoisseurship in the 17th and 18th centuries. He is currently working at the Eberhard Karls University Tübingen.

**Patrick Jentsch** is a Computer Scientists working in the interdisciplinary field of Digital History at Bielefeld University developing tools and workflows for computational text analysis.

**Anne Lappert** specializes in the field British Literary and Cultural Studies at Bielefeld University working on the 18th century British novel within the SFB 1288 “Practices of Comparing” at Bielefeld University.

**Malte Lorenzen** holds a PhD in German Studies specializing in journal studies, reception studies, and German literature and culture in the early 20th century at Bielefeld University.

**Anna Maria Neubert** holds a Master's degree in Digital Humanities and is a Doctoral Researcher in the Collaborative Research Center SFB 1288 at Bielefeld University.

**Christine Peters** specializes in Literary Studies and especially the study of the works of Alexander von Humboldt and works as a PhD within the SFB 1288 "Practices of Comparison" at Bielefeld University.

**Stephan Porada** holds a Master of Science degree in Interdisciplinary Media Studies and works within the SFB 1288 "Practices of Comparing" at Bielefeld University developing tools and workflows for computational text analysis.

**Kirill Postoutenko** is a specialist in historical semantics, history of totalitarianism, media and communication history. He is currently working in the SFB 1288 "Practices of Comparing" at Bielefeld University.

**Olga Sabelfeld** specializes in the history of practices of comparison in the genre of parliamentarian debates in the 20th century with a focus on Conceptual History working within the SFB 1288 "Practices of Comparing" at Bielefeld University.

**Helene Schlicht** is junior librarian at the University Library Johann Christian Senckenberg in Frankfurt am Main where she works in the field of research-related services.

**Ralf Schneider** is Professor and Chair of English Literature at the Institute of English, American and Romance Studies (IfAAR) at RWTH Aachen University.

**Silke Schwandt** is professor for Digital History at Bielefeld University focusing on the digitization of practices in the Humanities.

**Willibald Steinmetz** is professor for Political History at Bielefeld University specializing in Historical Semantics as well as European History of the 19th and 20th centuries.



