Analysis of Textual Variants with Robust Machine Learning Methods: Towards Novel Insights for the Digital Humanities

vorgelegt von
M. Sc. David Lassner

an der Fakultät IV – Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
– Dr. rer. nat –

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender	Prof. Dr. Wojciech Samek
Gutachter	Prof. Dr. Klaus-Robert Müller
Gutachterin	Prof. Dr. Anne Baillot
Gutachterin	Prof. Dr. Christiane D. Fellbaum

Tag der wissenschaftlichen Aussprache: 14. Februar 2023

Berlin 2023

Abstract

The analysis of textual variants allows us to explore how a given literary text came into being. This involves the analysis of the author's writing process and aesthetic inspirations but also analysing others who influenced the history of the text, for example as a translator or as a publisher. This further involves the inquiry into the historical and social circumstances under which the text was carried forward. These investigations are particularly relevant to the humanities not only because important contributions can be attributed to the respective person, but also because together they are the basis on which literary scholars then constitute *the text*: By comparison they can decide which textual variant is most adequate in the context of their specific research.

The path digital editors take to approach this question is that they gather source material to identify traces of alterations and compile a document with a complex annotation structure that contains all available textual variants.

Machine learning and computational humanities methods have the potential to contribute to this type of research in several ways, because they can (1) improve data availability with automated enrichment, (2) examine a broader collection through their ability to process textual sources at high speed, and (3) expand the existing catalog of methodology in literary studies.

The major challenge is that methods in natural language processing as a sub-field of machine learning assume a simplified, linear textual basis, and thus are not able to compare different textual variants with each other.

In Part 1, it will be addressed how linear textual variants can be extracted from complex document structures so that existing text processing methods can be applied.

In Part 2, it will be investigated how the methodologies of the different disciplines of machine learning and literary studies can be connected, to ensure that the proposed method and the obtained findings present a useful contribution in the respective disciplines. Here, we focus on the notion of text representation and introduce the new Word2Vec with Structure Prediction method for generating text representations in the context of structured corpora and show how it benefits the digital humanities.

Finally in Part 3, novel, robust natural language processing methods that are capable of comparing different textual variants are presented and applied in two different research contexts: In the analysis of a historical collection of letters from individuals who shaped intellectual Berlin around 1800 and in the study of the famous Schlegel-Tieck Shakespeare translation, with its translatorship origin still partly unexplained today.

Overall, this work aims to illustrate how transdisciplinary research between literary studies and machine learning leads to new insights and thus benefits both fields.

Zusammenfassung

Die Analyse von Textvarianten ermöglicht es zu erkunden, wie ein vorliegender, literarischer Text entstanden ist. Dies umfasst die Analyse der künstlerischen Inspirationen, aber auch die konkreten Personen, die in ihrer Rolle, sei es bspw. Autor, Übersetzer oder Verleger, Einfluss auf die Textgeschichte hatten, bis hin zur Untersuchung der historisch-sozialen Umstände, unter denen der Text weitergegeben wurde. Diese Untersuchungen sind so relevant für die Geisteswissenschaften, weil sie gemeinsam Grundlage sind, auf der Literaturwissenschaftler dann einen adäquaten Text konstituieren. Der Ansatz der digitalen Editionswissenschaften ist es, dafür Quelldokumente sammeln, die Spuren von Textänderungen einer bestimmten Person bzw. eines bestimmten Kontexts enthalten und diese in einer komplexen Dokumentstruktur festzuhalten.
Methoden des maschinellen Lernens und der computergestützten Geisteswissenschaft haben das Potenzial auf verschiedene Weise einen Beitrag bei dieser Art der Forschung zu leisten, da sie mit automatisierter Anreicherung die Datenverfügbarkeit verbessern können, sie durch ihre hohe Geschwindigkeit im Verarbeiten von Textquellen eine breitere Quellensammlung untersuchen können und weil sie den existierenden Methodenkatalog der Literaturwissenschaften erweitern.
Die große Herausforderung besteht darin, dass existierende Textverarbeitungsmethoden (NLP Methoden) üblicherweise von einer vereinfachten, linearen Textgrundlage ausgehen, also nicht in der Lage sind, verschiedene Textvarianten (*textual variants*) miteinander zu vergleichen.
In dieser Arbeit wird deshalb in einem ersten Schritt erarbeitet, wie aus diesen komplexen Dokumentstrukturen lineare Textvarianten extrahiert werden können, sodass bestehende Textverarbeitungsmethoden angewandt werden können (Teil 1).
In einem zweiten Schritt wird erarbeitet, welche Scharnierstellen es gibt, die die Methodiken der stark unterschiedlichen Disziplinen des maschinellen Lernens und der Literaturwissenschaft verbinden, sodass sichergestellt ist, dass die entwickelten Methoden und die damit erzielten Erkenntnisse in der jeweiligen Disziplin auch verwendbar sind. Dabei wird der Fokus auf den Begriff der Textrepräsentation gelegt und die neue Methode 'Word2Vec with Structure Prediction' zur Erzeugung von Textrepräsentation im Kontext von strukturierten Korpora vorgestellt und gezeigt, wie diese in den digitalen Geisteswissenschaften verwendet werden kann (Teil 2).
Zuletzt werden neue, robuste NLP Methoden vorgestellt, die in der Lage sind, Textvarianten zu vergleichen und diese werden in zwei verschiedenen Forschungskontexten angewandt: Bei der Analyse einer historischen Briefsammlung von Personen, die das intellektuelle Berlin um 1800 geprägt haben und bei der Untersuchung der berühmten Schlegel-Tieckschen Shakespeareübersetzung, mit ihrer bis heute teils ungeklärten Übersetzungsurheberschaft (Teil 3).
Insgesamt soll diese Arbeit verdeutlichen, wie eine transdisziplinäre Forschung zwischen Literaturwissenschaft und maschinellem Lernen produktiv neue Ergebnisse in beiden Feldern liefern kann.

Acknowledgements

This dissertation journey started with a coincidence. If Klaus hadn't known Anne personally, and if he hadn't put up a meeting with the three of us after I asked him if I could work on a master thesis on the application of machine learning on literary text – I hadn't heard of 'digital humanities' yet – I am certain things would have turned out very differently.

I want to express my deepest gratitude to both of my supervisors, Anne Baillot and Klaus-Robert Müller. Thank you Anne for letting your enthusiasm for scholarship spark over to me. For sharing your expertise on and guidance into the field of digital humanities, for encouraging me to deeply engage with the humanities and also for introducing me to the wonderful DH community.
Thank you Klaus for giving me the freedom to pursue my interests and for the guidance in pivotal moments. It was a pure joy to be part of the lab that you have created and that is truly a very special. I could not imagine any better place for being a doctoral student.

In recent years, I have had the pleasure of working with many colleagues, and I would like to mention a few of them in particular.
Thanks to Andreas Ziehe, Felix Biessmann, Julius Coburger, Kristof Schütt, Lukas Muttenthaler, Maximilian Alber, Miriam Hägele, Oliver Eberle, Philipp Seegerer, Seul-Ki Yeom and Thomas Schnake. Thank you Shinichi Nakajima for teaching me how to navigate reviewer comments and for often wearing the hat of the most critical reviewer yourself and thereby challenging me to the most principled solutions.
Thank you Sergej Dogadov for spending hours in front of whiteboards and passionately! discuss the mathematical details of our methods. How much joy you gave me whenever you made a discovery and you were too excited to tell that you were at the same time annoyed that speech is so slow and you couldn't present it faster.
I am also deeply thankful for the collaboration with Stephanie Brandl who I admire especially for teaching me how to retain an outside perspective and the big picture of how it all fits together even when being super involved with the research and at the same time.
I would also like to mention the three wonderful colleagues Christopher J. Anders, Jacob R. Kauffmann and Malte Esders who shared the office with me. Working among you felt like a dream come true, to share any thougts with you and discuss novelties in the field (or any recent hickups on the compute cluster) over the countless lunches and coffees we had together. Thank you.

I want to thank the fantastic people who all had an influence on my dissertation work in their own way: Andy Janco and the whole New Languages for NLP team, and Laurent Romary and Clemens Neudecker for their help with TEI and OCR, respectively.

I would also like to thank Christiane Fellbaum for accepting my invitation and taking the time to be a member of my doctoral committee, I am very honored.

In addition, I would like to thank Thomas Lassner and Emanuele Sbardella and, again, Stephanie Brandl for their feedback on the manuscript.

Finally, I am deeply thankful for my family, first and foremost my wife, Maike, who accompanied me on every step of this journey.

Contents

Chapter 1

Introduction

A major aspect of textual scholarship in the humanities is the analysis of textual variants. Scholars in the digital humanities (DH) encode the different variants in a complex, non-linear document structure. On the other hand, natural language processing (NLP) methods use simplified, linear text representations and without adaption cannot be used to analyse complex document structures. By extracting one, arbitrary, linear textual variant from the documents, NLP methods could be used for the analysis, however, in many settings this approach is not sufficient, but it is in fact needed to compare different textual variants.

As the analysis of textual variants is important in many sub-fields in the humanities, enabling NLP tools to process these types of complex documents has the potential for high impact and close collaboration between the two fields of science. In this thesis, three different digital humanities research settings are presented that demonstrate the importance of document structures and how they can be leveraged by NLP methods. One is the analysis of the writing and production of text (text genetics): In this setting, scholars acquire and compile evidence of the writing process of an author, mainly manuscripts that show alterations, comments etc. (Ehrmann 2016; Baillot and Schnöpf 2015). Understanding the writing process is a fundamental question in literary studies as it can highlight, for example where an author took inspiration from or who else was involved in the writing process. – As each edit can be seen as a branch from the otherwise linear textual document, describing the resulting document structure as a tree is more suitable.

A second setting is the analysis of translator style. In this case, scholars try to identify the style of a translator and distinguish it from the styles of other translators (Caballero, Calvo, and Batyrshin 2021; Rybicki and Heydel 2013; J. F. Burrows 2002). This type of research is especially important when the translation setting of a work is unclear and the contributions of each translator has to be identified for acknowledgement. When analysing different translations of the same original document, all translations, again, can be seen as textual transformations from the original, hence forms of textual variants. A stylometric analysis is thus much more profound when the variants are not considered in isolation.

A third setting is the syntactic and semantic dynamics of words across structured corpora, as comparative efforts in the literary studies are typically subject to struc-

tural criteria, such as author or genre (Calvo Tello 2021; Underwood 2014).
The goal of this thesis is to develop NLP methods that can be used to analyse textual variants in various settings in the digital humanities, by first making existing NLP methods to work with complex documents, by then aligning the methodologies of traditional literary studies with possible interventions from machine learning (ML), and by ultimately developing new NLP methods that address the challenges of highly structured documents and corpora thereby enabling the investigation of unanswered questions for the Digital Humanities.

1.1 Contributions and Outline of the Thesis

As a cross-disciplinary effort, this thesis is aimed at researchers from the ML and NLP community, as well as scholars from the DH. While it is our goal to deeply engage with all communities throughout the thesis, some chapters focus on a specific perspective. That means that while the thesis is divided into three abstract parts, (1) connecting DH, NLP and ML, (2) representing textual data and (3) modeling humanities phenomena, the chapters within are dedicated to a specific aspect or research question.

Part 1 (Enabling Cross-Disciplinary Resarch): This part demonstrates the relevant steps in order to apply NLP methods on DH data sets. The main contribution is the development of a software package that is capable of converting arbitrary textual variants from complex documents (Chapter 2). With this, scholars are able to extract a specific textual variant from their complex document structure and thereby have the chance to apply on it virtually any standard NLP method.
In Chapter 3, the opposite path is taken: machine learning is not used to analyse a human-annotated corpus but it is used to enrich a corpus that hasn't been annotated so far. Specifically, a corpus of Shakespeare translations from around 1830 that only existed as scanned images of the prints is automatically transcribed to allow for further analysis. In the process, a novel data set is created to improve optical character recognition (OCR) on historical prints from a specific period, more generally. It is shown how such data set can be published into the community for reuse, even in the presence of possibly copyrighted material. Overall, in this part the practical groundwork is laid for the rest of the thesis as the availability of data and software to conduct planned research is clarified and it is shown how it is possible for researchers of machine learning to either interact with complex DH documents fruitfully or enrich existing documents of lesser annotation complexity to increase their value for DH scholars.

Part 2 (Representing Textual Data): This part is about the theoretical and methodological prerequisites to conduct meaningful DH research using machine learning methods.

After introducing the fundamental concepts of text representations and transformations of textual representations from the NLP/ML perspective (Section 4.1) and likewise from the perspective of (computational) literary studies (Section 4.2), it is discussed in Section 4.3 how the two very different views can be connected: This is done by comparing the processes of transformations done by a human literary critic and ones that are machine learning based to see how methods can complement each other, discussing how certain parts of a literary critic's methods could be operationalized by machine learning representations and transformations.
By being orders of magnitudes faster than humans in processing text, ML methods have the potential to open up the canon of traditional literary studies, while at the same time bearing the risk of continuing or even amplifying biases present in traditional selection criteria of literary studies.
In Chapter 5 the novel Word2Vec with Structure Prediction method is introduced that yields word representations in the context of structured corpora. It is shown how this method can be used to compare different author's works with each other and investigate how the found similarities are aligned with known properties of the authors' works, such as genre, date of publication and location of writing.

Part 3 (Modeling Humanities Phenomena): In this part, two case studies are presented for concrete DH research projects. In Chapter 6, letters from around 1800 are investigated as a dynamic structure. The letters contain alterations stemming from the writers of the letters, the recipients, other persons the letters circulated to, but also archivists or descendants who were modifying the letters decades later for various reasons. The letters are therefore not documents of plain text but have a complex editing history. The alterations were analyzed with a modified LDA method, the novel AlterLDA. This analysis was able to confirm the presence of certain reasons for alterations but was also able to guide the attention to parts of the corpus that haven't been analyzed to that end before.
In Chapter 7, results on the analysis of the Schlegel-Tieck Shakespeare translation are presented, a corpus of the famous Shakespeare translation into German from the early 19th century that was created as a collaborative effort by three different translators. The study analyses the style of the plays where the true translator is unknown based on the style of the part of the corpus where the translator is known. It considers both, the source text and the target text for the stylistic analysis of the translators. Apart from insights into the collaborative translation setting, this research underscores the important role that Dorothea Tieck played in the effort, a role that she was not acknowledged for in the translated publication at that time.

1.2 Included Papers

The following six publications are the main building blocks of this dissertation. The co-authors thankfully agreed to me taking these papers and reuse or adapt them to appear in this dissertation.

1. David Lassner, Julius Coburger, et al. (2021). "Publishing an OCR ground truth data set for reuse in an unclear copyright setting". In: *Zeitschrift für digitale Geisteswissenschaften* Sonderband 5, Fabrikation von Erkenntnis – Experimente in den Digital Humanities, online. DOI: 10.17175/sb005_006

2. David Lassner. "The Standoff Converter - Bridging the gap between NLP and TEI" under review

3. Anne Baillot and David Lassner (2022). "Von Graphen zu Word Embeddings. Zur Entwicklung des mathematischen und visuellen Instrumentariums der Literaturwissenschaft". In: *Germanica* 71 (2), pp. 191–203

4. David Lassner, Stephanie Brandl, et al. (2023). "Domain-Specific Word Embeddings with Structure Prediction". In: *Transactions of the Association for Computational Linguistics* 11, pp. 320–335

5. David Lassner, Anne Baillot, Sergej Dogadov, et al. (2021). "Automatic Identification of Types of Alterations in Historical Manuscripts". In: *Digital Humanities Quarterly* 15.2, online. URL: http://www.digitalhumanities.org/dhq/vol/15/2/000553/000553.html

6. David Lassner, Anne Baillot, and Julius Coburger (2019). "Attributions Of Early German Shakespeare Translations". In: *Book of Abstracts of the Digital Humanities Conference*. DOI: 10.34894/DK6QKN

Part I

Enabling Cross-Disciplinary Research

Chapter 2

Extracting Textual Variants from Complex Documents

There is a fundamental difference between the way digital philologists work with their documents and the way NLP scholars work with their data. In NLP, there exist corpora which are 'just plain text' but also some corpora that have extensive word-by-word annotations, for example for morphological linguistic analysis (Nivre et al. 2020, as a famous example). Instead of plain text, digital philologists often create genetic editions and annotate corpora word-by-word or even character-by-character.

Fine-grained annotation is costly because it involves human labour and in many cases even labour of specialists in their fields, let it be linguistics or literary scholarship.

The corpus a literary scholar works on, can be seen much more as a repository of sources and annotations of sources (the apparatus). Often, it is a collective effort of multiple scholars each pursuing their own research questions. One might be diligently annotating each little textual variant in a manuscript (focus on the document), another one might be interested in identifying a coherent text as it was intended by the author to be read (focus on the text).

In order to apply an NLP method on the corpus and to have it return meaningful predictions for the DH scholar, it has to be given an appropriate, well prepared textual variant. In this chapter, the fundamentals of text encoding in the digital humanities and digital genetic editing are introduced. Then, the Standoff Converter is presented, a software package with which a scholar can extract a specific textual variant from a genetic edition by making explicit decisions about filtering and enhancing the textual sources.

2.1 Fundamentals of Text Encoding in DH and NLP

In DH, the de facto standard for encoding text is the text encoding initiative (TEI).[1] TEI offers guidelines on how to embed textual documents in XML adding meta data and annotations. A typical TEI-XML document structure is given in Listing 2.1.

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <TEI xmlns="http://www.tei-c.org/ns/1.0">
3     <teiHeader>
4        <fileDesc>
5           <titleStmt>
6              <title>...</title>
7              <author>...</author>
8              <editor>...</editor>
9           </titleStmt>
10          <publicationStmt>
11             <authority>...</authority>
12             <availability>
13                <licence target="..."/>
14             </availability>
15          </publicationStmt>
16          <sourceDesc>...</sourceDesc>
17       </fileDesc>
18    </teiHeader>
19    <text>
20       <body>
21          ... <persName>...</persName> ...
22          <note>...</note> ...
23       </body>
24    </text>
25 </TEI>
```

Listing 2.1: Typical TEI-XML document skeleton with examples for meta data such as author, title and editor and for textual data with persons annotated as named entities.

It comprises a header `teiHeader` with all general meta data regarding the TEI file (licenses etc.) but it also includes information on the underlying source (material, author, etc.). Then, there is a `text` part that contains the transcription of the source text but also inline annotations such as named entities (`persName` etc.) or notes. The notes could be footnotes of the source document but could also be notes by the editor of the TEI document. This may be the place for a critical apparatus of a digital scholarly edition.

[1] See `https://tei-c.org`.

Evidently, the TEI document can be much more complex than a plain text transcription of a source document. Most importantly, there are different potential plain texts that can be extracted from the TEI document. A simple example would be to either extracting the plain text including the critical apparatus or extracting the plain text excluding the critical apparatus. Another example would be hyphens: when a textual document is digitized it will usually contain hyphens at the end of text lines to break up a word and to make the reader aware that the word continues on the next text line. With standard optical character recognition (OCR) the punctuation character (usually '-') is simply preserved as a text symbol. Most likely a naïve digital rendering of the document would not have the same horizontal space as the source page which would result in hyphenation characters scattered all over the text. Alternatively, the punctuation characters could be removed but then it is not possible to reconstruct the original line length that might be very important in certain research settings (when analysing meter, for example). This comes in addition to the question of whether a line break is a breaking one (e.g. a single space should be inserted) or a non-breaking line break (e.g. no single space should be inserted because the word continues on the following line). In TEI, both issues can be addressed by explicitly annotating the hyphenation and the line break, for example in Listing 2.2.

```
1 ...
2 <text>This is a sent<pc>-</pc><lb break="no" />ence with a line break
     .</text>
3 ...
```

Listing 2.2: An example sentence encoded in TEI with a line break and hyphenation encoded with 'pc' followed by the line break. that specifies that the word should not be split by broken in two.

Where `pc` stands for punctuation character and `lb` stands for line break.
This means that a TEI document offers both versions for different research questions where in a plain text document this decision has already been made and certain research questions are impossible to pursue.
At the same time the TEI document combines the representation of the source with the research output: If, for example, a digital philologist would prepare a critical digital edition of a work then their apparatus would live in the TEI document embedded in the source transcription. The TEI document would then contain annotations specific to the research question of that digital philologist. – Where this strategy (in comparison to separating transcription and annotation into different documents) has clear advantages for longevity it may be at cost of interoperability between different digital scholarly editions to that extend that there are often edition-specific annotations.

2.2 Genetic Editing with TEI

Classical scholarly editing has a long-standing tradition in distinguishing between different types of editions (Witkowski 1924). The characteristics of specific edition forms usually align with the intended readership, but they also take into account a bibliographic history that tends to differentiate more and more along time according to linguistic areas. In the German-speaking area, historical-critical editions that comprise an extensive historical-critical apparatus are often distinguished – with a clear hierarchical difference – from so-called study editions (Plachta 2006). The common denominator between these two types of editions is that they aim to offer a "reliable text" as a central component (Plachta 2006). In contrast to these types of editions, it is also possible to publish a reproduction of the manuscript image (facsimile edition). Plachta points out, however, that a facsimile edition is no substitute for the above two types of editions (Plachta 2006).

Another way of differentiating between types of editions is to compare the intention in the text construction, which corresponds to the philosophy according to which the anglo-saxon area has mainly structured their approach. According to Andrews, "the 'old' methods that have their root in classical philology" strive to assemble the "ideal" text, while the "new philology" seeks to find the "real" text (Andrews 2013). In this conception, the ideal text tries to approach the author's intention, while the real text seeks to emulate the existing sources.

The type of edition an editor goes for is often defined by economic factors in printed editions, while in digital editions, this limitation can be obsolete in terms of the amount of pages available, or located on a different level (for instance due to the price of specific, cost-expensive technologies). More generally, in digital scholarly editions, differentiation characteristics can be renegotiated. As Andrews states, there are hardly any technical limitations in digital editions with regard to the size of the apparatus, and the number and resolution of facsimiles provided (Andrews 2013).

This is not the only specificity that distinguishes digital from print editions. They also are machine-readable. With digital editions being available in digital formats, computers can not only handle repetitive tasks in the creation of the edition (Andrews 2013), they can also be used to perform tasks that use the edition as source material. The most obvious example for this type of use is the full-text search, but the machine-readable form also allows the creation of a multitude of statistics and customed visualizations with very little effort (Ralle 2016). Furthermore, Ralle emphasizes that the digitization of editions and scholarly editing in general allow to pay special attention to the processual aspects of the edition (Ralle 2016). An edition can be extended or enriched after it has been initially published and does not need to be 'finished' at a specific moment in time. A digital edition can be modified dynamically, for instance like the Carl-Maria-von-Weber-Gesamtausgabe with a front page field called 'What happened today?' that connects to all instances of the current date in the corpus and highlights them – a content that changes from day to day and offers a different approach to the corpus than the traditional keyword search. Also,

user interaction can be funneled back into the edition, for example when subsequent publications that are based on the edition are listed there. Interaction in and of itself can also be included: the search behaviour of users can be analyzed for better future suggestions or the edition can be enriched by third-party data. Every user of a digital edition, whether computer or human, is thus potentially able to engage in one form of editorial participation or the other (Schlitz 2014; Siemens et al. 2012; Shillingsburg 2014).
These special features of digital editions allow for paleography (Baillot and Schnöpf 2015) to reach out to research questions hence unexplored in the humanities due to the lack of tools and corpora allowing an automatic evaluation of alteration phenomena. It enables for instance to thoroughly reconstruct the history of a document by considering physical traces of alterations, meaning any smaller or larger text modifications on the manuscript, performed either by the author himself or herself or by others. This approach provides insights into the way in which authors, editors and other contributors work together, hence impacting our understanding of text genesis as a collaborative process.
In order to achieve substantial results in this field of research, fast and well-structured access to the document variants is required. Digital editions presenting the manuscript alterations allow to focus on diplomatic transcription or facsimile, as opposed to print editions where the focus is on a single copy text, itself usually optimized for readability. Examples of digital editions representing the document history include faustedition.net (Goethe 2022) , bovary.fr (Flaubert 2009), beckettarchive.org (Beckett 2022), and the edition that will be the focus or analysis in Chapter 6, the: the digital scholarly edition "Letters and texts. Intellectual Berlin around 1800", `berliner-intellektuelle.eu` (Baillot 2022), BI in the following.

2.3 The Standoff Converter

In the previous sections the importance for TEI in the digital philologists community has been shown also the state of data set standards for NLP has been sketched out briefly. The first question is: if TEI is the superior encoding why isn't it used more broadly outside the digital philologists community?
The fact that TEI has the capabilities to encode different potential plain texts implies that whenever a specific plain text is needed in the context of a research project, this specific plain text has to be extracted from the TEI. This is an additional preprocessing step that needs to be performed – additional configurability of the source data comes with the cost of additional work on preprocessing the source data that the NLP community might have been negligent upon in the past.
The second question is: if there is such a growth of accuracy of NLP models on various tasks such as linguistic annotation or named entity recognition in recent years, why did these automatic annotations not yet end up in TEI editions more broadly? Again, as for the first question, the TEI itself usually cannot be passed into an NLP model directly, instead, a plain text version has to be extracted, but more impor-

tantly, the predictions of an NLP model that received the plain text as input will also only apply to that plain text. So what is missing here is the procedure to add the predictions of the NLP model as annotations to the original TEI document.
In order to simplify this process, the so-called Standoff Converter was developed. The Standoff Converter is a Python package that offers interaction with TEI documents programmatically. It can switch between an lxml[2] tree view, a standoff view and a plain text view and therefore bridges the gap between digital philologists, who often work with TEI, and NLP scholars, who mostly work with plain text. It was developed to address a well-identified need in the methodological interaction between NLP and Digital Humanists working with structured data. The Standoff Converter is open sourced (`https://github.com/standoff-nlp/standoffconverter`), it is documented (`https://standoffconverter.readthedocs.io/`) and it's development is test-driven.
To illustrate how the Standoff Converter functions, we will present a simple toy example that is given in Listing 2.3.[3]

```
<TEI>
<teiHeader> </teiHeader>
<text>
    <body>
        <p>1 2 3 4. 5 6<lb/> 7 9 10.</p>
        <p> 11 12 13 14</p>
    </body>
</text>
</TEI>
```

Listing 2.3: A simple TEI document that is used throughout this section to illustrate how the Standoff Converter works. It shows three sentences in two different paragraphs and it has a line break annotated.

This TEI document contains two paragraphs with three sentences in total. Also, one can see that one line break is encoded with an `<lb/>`-tag. In this example, we consider the task to split the document into sentences and annotate the sentences with sentence tags.
To process the file, it is first parsed as an lxml etree which is a standard way to deal with XML in Python. Then, a Standoff object is created that separates the text from the annotations, shown in Listing 2.4.[4] This `Standoff` object is part of the Standoff Converter package, hence our contribution.

```
tree = etree.fromstring(input_xml)
so = Standoff(tree)
```

[2]Python library for processing XML documents - `https://lxml.de`.

[3]The same example was also used in the interactive web demo that can be visited at `https://so.davidlassner.com/`.

[4]The idea of separating text and annotations has already been discussed (and advocated for) in the context of the Digital Humanities, see Schmidt 2016.

Index	Context	Text
0	text	
1	text>body	
2	text>body>p	1 2 3 4. 5 6
3	text>body>p>lb	
4	text>body>p	7 9 10.
5	text>body	
6	text>body>p	11 12 13 14
7	text>body	
8	text	

Table 2.1: Example of the collapsed table view.

```
3 print(so.collapsed_table)
```

Listing 2.4: Two-step conversion from input xml to an lxml tree and to a standoff object.

The output is shown in Table 2.1. The Standoff object contains a table where for each character the corresponding context is stored. Here, we can see the so-called collapsed table, which shows each context and all characters that have the same context merged into one row for readability. This uncluttered view on the document can be immensely useful when exploring unknown TEI documents. It often makes the text more readable while at the same time you can see where you are in the document.
The text column in this representation contains all text of the `<text>` tag of the TEI document including the code indentation and line breaks. Of course it does not contain a line break at the `<lb/>` position. (2) The next step is to create a view of the text that includes all the parts that are relevant for our NLP task. This View object is also part of the Standoff Converter package. In this case, we insert a line break for the `<lb>` tag and exclude all text outside the `<p>` elements.

```
1  view = (
2      View(so.table)
3          .insert_tag_text(
4              "lb",
5              "\n"
6          )
7          .exclude_outside("p")
8  )
9
10 plain = view.get_plain()
11 print(plain)
12 > 1 2 3 4. 5 6
```

```
> 7 9 10. 11 12 13 14
```

Listing 2.5: Example of the callable query style of the plain text view that, in this case, replaces the line break tag with a libe break character and that discards white space outside the paragraphs (the indentations).

The output is shown in Listing 2.5. (3) Next, one can apply an NLP tool on the plain text, in this case spaCy's standard English sentencizer (Honnibal et al. 2020). The found sentences are depicted in Listing 2.6.

```
from spacy.lang.en import English
nlp = English()
nlp.add_pipe('sentencizer')

sentences = []
for sent in nlp(plain).sents:
    print(f"* {sent}")
    sentences.append(sent)
> 1 2 3 4.
> 5 6 7 9 10.
> 11 12 13 14
```

Listing 2.6: Example of the found sentences in the plain text view.

(4) Finally, the found sentence boundaries are translated into standoff character positions and new annotations are added to the TEI. Crucially, the View object retains a reverse lookup that can translate positions in the plain text string (`sent.start_char`) into indices in the standoff table `start_ind`. Parameterized with the indices, new annotations can be added to the standoff table (`add_inline`) that are immediately synchronized with the tree data structure `lxml.etree`, as shown in Listing 2.7.

```
for isent, sent in enumerate(sentences):

    start_ind = view.get_pos(sent.start_char)
    end_ind = view.get_pos(sent.end_char-1)+1

    so.add_inline(
        begin=start_ind,
        end=end_ind,
        tag="s",
        depth=None,
        attrib={'id':f'{isent}'}
    )
```

Listing 2.7: Character positions are converted to standoff table positions and new annotations are added to the standoff table.

Afterwards, the final TEI document contains sentence tags, as shown in Listing 2.8.

```
<TEI>
<teiHeader> </teiHeader>
<text>
   <body>
      <p><s id="0">1 2 3 4.</s> <s id="1">5 6<lb/> 7 9 10.</s></p>
      <p> <s id="2">11 12 13 14</s></p>
   </body>
</text></TEI>
```

Listing 2.8: Final output of the TEI document after the new sentence tags were added by the Standoff Converter.

This simple example was meant to illustrate how the package can be used to extract a specific textual variant from a TEI document and add new annotations. In a larger research project, one might want to separate manual annotations and automatic annotations to the document. Further details about the software architecture design are given in Appendix A.

2.3.1 Contributions to the TEI community and applications of the Standoff Converter

This section highlights in what ways the Standoff Converter contributes to the TEI community. The Standoff Converter is a novel Python package that makes it possible to apply NLP models on TEI documents. It is based on the lxml package that is the standard way of parsing XML in Python - but unlike lxml, the Standoff Converter is TEI-aware. As it presents a table-based view of the TEI document, in addition to the tree-based view of lxml, it is more accessible to learners of TEI.
The TEI Guidelines state that the new standOff element should include "[..] content that does not fit well in the text" but rather contextual information. Another important point is that, typically, digital philologists are also unwilling to have automatic annotations mixed up with their carefully crafted digital edition. There are indeed good reasons for this: Automatic annotations might be inaccurate, and they may also blow up the amount of annotations within the `<text>` element. To make a clear distinction, one can add all automatic annotations to the standOff element. Then, manual annotations and automatic annotations are separated from each other. This pathway to organize the TEI document would be a useful addition to the TEI Guidelines. The Standoff Converter can be of help in this case, as one can freely decide where annotations should be added or moved to: inline or standOff.
Banski et al. 2016 identified that besides further development of the Guidelines, the one thing that has the potential to improve the support of standoff in TEI would actually be better guidance of how to use the existing support. The Standoff Converter - as a tool - can help with such guidance by implementing sensible default behavior. There are two examples where guidance is needed. (1) One can either use annotation triples (body - annotation - target) or one can use annotation tuples.

The triplet approach is in line with the Open Annotation Data Model[5] but in this model, multiple features should be added to the same annotation. Whenever a new annotation is added to the standOff element, one has to look for existing similar spans. With the tuple approach one can simply add a new span-feature combination each time.
(2) Either ids or string-ranges can be used to reference the text that is annotated in the standOff element. The internal data structure of the Standoff Converter corresponds naturally with using the string-range pointer approach in the standOff element. The most direct conversion from the Standoff Converter's data structure to string-range-based standOff elements would be to reference all spans relative to the `<text>` element. The other extreme would be, as also discussed by Banski et al. 2016, to wrap each word with an id'ed <w> tag and reference every word by id.
Both approaches have their own disadvantages, because the former one would likely invalidate when any text inside the `<text>` element changes and the latter would blow up the overhead of annotation within the `<text>` element.
Aiming for a middle ground of adding inline sentence tags and using string-range pointers relative to the sentence tags (that one can be added automatically as shown in the showcase earlier) might be a good starting point. This way, the overhead of inline annotations is acceptable and the string-range pointers in the standOff element only break when the corresponding sentence is changed. This approach is typically a sensible default that could work for many TEI editions that plan to work with automatic annotations.
This package was conceived and developed to help scholars with their TEI projects and centralize the community efforts so that fewer ad-hoc solutions are created. In that line, the the work on building a community maintaining the Standoff Converter package has to be continued. Ultimately, the goals is that this package helps bridge the gap between the digital philology community and the NLP community in a way that is beneficial for both worlds.

2.4 Summary & Conclusion

In this chapter, the different scholarly practices between the NLP community and the DH community were presented with a focus on TEI as an encoding that is very popular among the DH community. To that end we have discussed the difference between a textual document (digital or not) and 'the text' and that it, amongst others, depends a lot on the framing, target group, type of edition, research question and the scholarly tradition what of the document is actually considered 'the text'. The Standoff Converter was proposed, a Python package that simplifies the technical aspects of extracting a plain text version from the TEI document and that retains a back link that enables the creation of annotations in the original TEI document from predictions of an NLP model on the plain text.
We have shown in what ways NLP scholars could benefit from using TEI and how

[5] See `http://www.openannotation.org/spec/core/`.

digital philologists integrate NLP into their editions by using the Standoff Converter. The use of the Sandoff Converter therefore improves the workflow for both research communities.

With regard to the Standoff Converter as an open source software project, it turned out that writing the software and adopting good software development practices such as test-driven development and documentation is only partly what is needed to get attention from other scholars and ultimately found a community that helps maintain the project – additionally a large portion of the work is actually community building and making sure the project gets attention repeatedly in academic social media.

Apart from this long-term task, there are other directions toward which the Standoff Converter can develop: an obvious continuation would be to better integrate with the `standOff` element of the TEI so that, for example, the Standoff Converter can directly write annotations into the `standOff` element and also interchangeably switch annotations between inline and standoff in the TEI document tree.

Another future direction would be to offer a TEI visualisation by the Standoff Converter, by converting parts of the TEI into a scalable vector graphics (SVG). Similar to a plain text view an SVG view could be created that could retain certain layout properties from the TEI document but at the same time wouldn't have to be as flat as plain text. There exist solutions for rendering OCR ground truth (like PAGE or ALTO)[6] as SVG in eScriptorium (Kiessling et al. 2019) and even rendering TEI as SVG in the teipublisher[7] and it would certainly be fruitful to integrate the existing rendering solutions as part of the SVG-view in the Standoff Converter.

[6]For ALTO, see the description of the standard at the Library of Congress `www.loc.gov/standards/alto` and for PAGE, see Pletschacher and Antonacopoulos 2010.

[7]See `https://teipublisher.com`.

Chapter 3

Enriching and Publishing Humanities Data with Machine Learning

In Chapter 2, the proposed way to connect NLP and DH was by acknowledging the complexity of DH corpora and documents, and showing solutions how to still make existing NLP tools work that have originally been designed for less complex document structures.
The basis for this kind of research is that such elaborately annotated corpora exist. Unfortunately, this is only true for a very small part of literary history, although from today's perspective other parts (for example other authors than the ones belonging to the literary canon) would deserve to be studied as well. Therefore, this chapter addresses the complementary question, namely how ML methods can work with the available material of worse annotation quality to contribute to DH research questions, and whether ML methods can also help to make new corpora accessible and enriched. In the scarce data setting, ever so often there are both ways to develop a research question, by either specifying the data that should be included or by leveraging what data is actually available – and especially in what form it is available.
There are different forms in which digital textual material is available and, trivially, the amount of available data is anti-correlated with the 'quality' of data. The term quality is not yet strictly operationalised but as an example, a digitized historical book can be published as one of the following:

- Scan: a collection of scanned page images.
- OCR: a document with a scanned image for each page and an automatically extracted text layer. This is the form, most books on Google Books are published.
- Corrected OCR/transcribed text layer.
- Added meta data/data sheet to the PDF.

- Added annotations such as NER/L or linguistic annotations automatically.
- Added/corrected annotations such as NER or linguistic annotations manually.
- Added other in-depth annotations or critical apparatus.

The different stages are increasingly costly and the 'quality' could be measured by the amount of preparation that went into the given publication. Again, there are, for example, millions of books available in Google Books[1] as a PDF of scanned images with a low-quality OCR'ed text layer. Corpora with high quality text layers (usually meaning manually created or corrected) are much smaller. As discussed previously, sophisticated digital genetic edtions are even rarer.

In NLP as well as in DH, new insights are often found when either new data becomes available or a method is found that can make use of (lower quality) data that could not be used before. For the latter, three examples will be given from NLP or DH research: Clearly, the emergence of large language models (LLMs) (or sometimes also referred to as foundation models) is the first example that comes to mind. large language models can be trained on large amounts of unlabelled text and are able to acquire detailed properties of language such that they can then be fine-tuned on another data set – the data set of interest.

> A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks.[2] (Bommasani et al. 2021)

Another example is the shift from character-level OCR to line-level OCR. Originally, OCR ground truth was created by annotating individual characters on a page with their transcription. But this was a very tedious process because it meant that the annotator had to create bounding boxes for each character. With Liwicki et al. 2007, a novel method was introduced that allowed to train a recognition model on line-level transcriptions with the help of a CTC loss. Line-level transcriptions are a lot less costly to produce and therefore ground truth data sets could be acquired much more cost efficient.
A third example is the analysis textual documents without OCR: the idea to analyse the scanned images directly. This can be of help if the documents have a complex layout or the font (or handwriting) is too irregular so that it is still very costly to create sufficient amounts of ground truth data. A specific similarity measure on the page images has been used to explore knowledge dissemination in historical astronomy books (see El-Hajj et al. 2022; Eberle et al. 2022). This stresses the point that it is as important to publish new data sets, as it is to use existing data efficiently and to republish any additions or enhancements for future research.
At the same time, when shifting away from small hand-curated, well documented

[1] A description of the process how the Google Books team processed the CHI's book collections can be found in Michel et al. 2011.

[2] In this thesis we refer to foundation models as large language models.

data sets toward larger and messier data sets, scholars might not be aware of certain biases within the data sets. Digital philologsts observe that the performance of NLP models strongly increased in recent years, for example for the task of Named Entity Recognition (Akbik, Blythe, and Vollgraf 2018; Li et al. 2020; Strubell et al. 2017; Yu, Bohnet, and Poesio 2020, to name a few). And leveraging the advances in NLP to add automatic annotations to their existing digital editions is a valid strategy and a way to bridge the gap between the two fields from the digital philologists perspective – In a way, accepting the uncertainty and imperfection of NLP models or messy data. From the perspective of NLP researchers, there has been a recent shift of focus toward more ethical considerations and the mitigation of biases in the underlying data sets. It has been proposed that there is a need for "[d]ata sheets for data sets" (Gebru et al. 2021) that would standardize the documentation of data sets for NLP. And in general, more resources are necessary related to curation and documentation of the underlying data sets used for large language models (Bender et al. 2021). – This shows that NLP researchers begin to acknowledge the incompleteness of data sources and the imperfection of categorization which has been fundamental part of digital philologists and archivists methodology.
In this chapter it will be shown how to collect a corpus of digitized PDFs, and how to create automatic transcriptions with ML methods. This way, machine learning helps with enriching digital corpora that can then be analyzed by scholars in more targeted ways.
It will be demonstrated how to evaluate the performance of the models and it will be asserted if it is sufficient for subsequent tasks.
Additionally, it is shown how the human-annotated training data, the models, as well as the automatically transcribed documents can be published in a way that they are easily reusable by other scholars to counteract the scarcity of digitally available historical and literary data as a community effort.

Digital access to cultural heritage has been improved by optical character recognition (OCR), which is the task by which a computer program extracts text from a digital image in order to draw the text from that image and present it in a machine-readable form. For historical prints, off-the-shelf OCR solutions often result in inaccurate readings. The results of an OCR method can be improved significantly by using a pre-trained model and fine-tuning it on only a few samples that display similar characteristics (Liebl and Burghardt 2020; Reul et al. 2017; Springmann et al. 2018). To that end, there has been a growing effort from the Digital Humanities community to create and publish data sets for specific historical periods, languages and typefaces aiming at enabling scholars to fine-tune OCR models for their collection of historical documents (Padilla et al. 2019).[3] In Germany, the DFG-funded OCR-D initiative brings together major research libraries with the goal to create

[3]For manuscripts, just recently the Transcriptiones platform launched, see Fuchs and Weber 2022. For French texts from the 18th to the 21st century there exists HTR-United, see Chagué and Clérice 2022. The slightly different approach of just publishing fine-tuned models for different settings is proposed by (1) Transkribus: `http://transkribus.eu/wiki/images/d/d6/Public_Models_in_Transkribus.pdf`, (2) READ-COOP

an open source framework for the OCR of historical printed documents, including specifications and guidelines for OCR ground truths (Engl et al. 2020).
In order to improve OCR results, images and the corresponding transcriptions are collected in such a way that each pair (image and text) only represents one line of text from the original page. This is called a ground truth data set and is precisely what we will focus on in the following.
Besides the fact that creating transcriptions of images manually is tedious work, another major issue arises from this type of collective effort in that the institutions that produce the scan often claim some form of copyright to it. For example, on the first page of any of their PDFs, Google Books "[…] request[s] that you use these files for personal, non-commercial purposes"[4]. As a consequence, a scholar aiming to create an OCR ground truth data set would not know with certainty whether the rights to redistribute the textline images derived from the PDF can be considered as granted.
The OCR ground-truth data set discussed in this chapter has an unclear copyright setting for the image data. The legal background is discussed, the relevance of the data set is shown and an in-depth analysis of its constitution and reuse is provided by investigating two different approaches to overcome the copyright issues.
In order to address these issues in the following two ways are compared to publish the OCR ground truth data set with image data.

- As Google Books works with cultural heritage institution (CHI) to digitize books, we asked permission from the CHIs to redistribute the image data.
- We published a data set formula, which consists of the transcriptions, links to the image sources, and a description on how to build the data set. For this process, we provide a fast, highly automated framework that enables others to reproduce the data set.

3.1 Legal Background and its Interpretation at CHIs

Clarifying the copyright situation for the scans of a book collection requires to take into account, for each book, the cultural heritage institution owning the book (usually a library), and, in the case of private-public partnerships, also the scanning institution (e. g. Google Books) involved in its digitization. For Google Books, there exist different contracts between CHIs and Google, and not all of them are open to public inspection. However, based on comparing the ones that are available, I assume that other contracts are to some extent similar (see List of Contracts). The contracts contain information on the 'Library Digital Copy' for which non-profit uses

2021: `https://readcoop.eu/transkribus/public-models/` and (3) (Kraken) OCR/HTR ocr_models: `https://zenodo.org/communities/ocr_models/`.

[4]Google Inc. 2006, cited after J. Ruiz 2011.

are defined under Section 4.8 (cf. British Library Google Contract), which states that a

> Library may provide all or any portion of the Library Digital Copy, that is [...] a Digital Copy of a Public Domain work to (a) academic institutions or research libraries, or (b) when requested by Library and agreed upon in writing by Google, other not-for-profit or government entities that are not providing search or hosting services substantially similar to those provided by Google. (British Library Google Books Agreement in J. Ruiz 2011)

When trying to unpack this legal information against the use case presented here, multiple questions arise. What are the legal possibilities for individual scholars regarding the use of the Library Digital Copy of a Public Domain work? How can there be limitations in the use of a Public Domain work? Is the use case of OCR model training substantially similar to any search or hosting services provided by Google? Would and can libraries act as brokers in negotiating written agreements about not-for-profit use with Google?
In the continuation of Section 4.8 of the contract, additional details are specified with regard to data redistribution by 'Additional institutions' where

> [a written agreement with Google] will prohibit such Additional institution from redistributing [...] portions of the Library Digital Copy to other entities (beyond providing or making content available to scholars and other users for educational or research purposes. (British Library Google Books Agreement in J. Ruiz 2011)

This brings up further questions but also opens the perspective a bit, since there appear to be exceptions for "scholars and other users for educational or research purposes"[5], which is a precise fit of the use case we present here. Now what does this mean in practice? Digital Humanities scholars are not necessarily legal experts, so how do libraries that have entered public-private-partnerships with Google for digitization of public domain works implement these constraints? Schöch et al. discuss a wide range of use cases in the area of text and data mining with copyright protected digitized documents, but they do not cover the creation and distribution of ground truth (Schöch et al. 2020).
In other scenarios that involve copyrighted texts published in derived formats, one question typically preventing redistribution is whether it is possible to re-create the (copyright-protected) work from the derived parts. In the case of textline ground truth, it is however likely that this would constitute a violation of such a principle. In this unclear setting, scholars are in need of support and guidance by CHIs.

We have asked ten CHIs for permission to publish image data that was digitized based on their collection in order to publish them as part of an OCR ground truth

[5] British Library Google Books Agreement in J. Ruiz 2011.

Institution	# books	# pages	Response time	Permission to publish	Permission to license	Alt. source	Responsible	Cit. needed
Bayerische Staatsbibliothek	4	12	3	yes	yes	yes	yes	yes
Biblioteca Statale Isontina Gorizia	1	3	–	–	–	–	–	–
Bodleian Library	11	20	2	yes, alt.	already CC	yes	yes	yes
British Library	1	35	4	no	no	no	yes	–
Harvard College Library	1	3	0	yes	yes	yes	no	yes
New York Public Library	5	29	3	–	–	no	no	no
Austrian National Library	2	6	10	yes, alt.	no	yes	yes	yes
Robarts – University of Toronto	2	3	–	–	–	–	–	–
University of Illinois Urbana-Champaign	6	4	0	yes	yes	no	yes	yes
University of Wisconsin – Madison	8	24	2	yes	yes	no	no	no

Table 3.1: Responses of library institutions to our request to grant permission to publish excerpts of the scans for which they were contractors of the digitization. Most institutions responded within a few working days and except for the fact that most acknowledged the public domain of the items, the responses were very diverse. Many answered that they are either not responsible or only responsible for their Library Copy of the PDF (Lassner, Coburger, et al. 2022).

data set under a CC-BY license. As shown in Table 3.1, the institutions gave a wide variety of responses. Many institutions acknowledged that the requested books are in the public domain because they were published before the year 1880. However, there is no general consensus on whether the CHIs are actually responsible for granting these rights, especially if one wants to use the copy from the Google Books or Internet Archive servers. Some institutions stated that they are only responsible for their library copy of the scan and granted permission to publish only from that source. Only two institutions, the Bayerische Staatsbibliothek and University of Illinois Urbana-Champaign stated that they are responsible and that we are allowed to also use the material that can be found on the Google Books or Internet Archive servers.

This case study underlines the lack of a clear and simple framework of reference that would be recognized and applied, and would reflect on good practices in the relationships between CHIs and digital scholarship. The lack of such a framework is addressed among others by the DARIAH initiative of the Heritage Data Reuse Charter[6] that was launched in 2017. Another approach towards such a framework is that of the 'digital data librarian' (Eclevia et al. 2019).

3.2 Description of the Data Set

In the data set that we want to publish in the context of our OCR ground truth, we do not own the copyright for the image data.[7] We therefore distinguish between the data set formula and the built data set. We publish the data set formula which contains the transcriptions, the links to the images and a recipe on how to build the data set.

The data set formula and source code are published on Github[8] and the version 1.1 We am referring to in this thesis is mirrored on the open access repository Zenodo.[9] The data set is published under a CC-BY 4.0 license and the source code is published under an Apache license.

3.2.1 Origin

The built data set contains images from editions of books by Walter Scott and William Shakespeare in the original English and in translations into German that were published around 1830.

The data set was created as part of a research project that investigates how to implement stylometric methods that are commonly used to analyze the style of authors with the goal of analyzing that of translators. The data set was organized in such a

[6] See Baillot, Mertens, and Romary 2016. For additional information on the DARIAH Heritage Data Reuse Charter, see data-re-use, *Cultural Heritage Data Reuse Charter* 2022.

[7] The current version of the data set can be found at `https://github.com/millawell/ocr-data/tree/master/data`.

[8] See `https://github.com/millawell/ocr-data/`

[9] See `https://doi.org/10.5281/zenodo.4742068`.

way that other variables like authors of the documents or publication date can be ruled out as a confounder of the translator style.
We found that 1830 Germany was especially suitable for the research setting we had in mind. Due to an increased readership in Germany around 1830, there was a growing demand in books. Translating foreign publications into German turned out to be particularly profitable because, at that time, there was no copyright regulation that would apply equally across German-speaking states. There was no general legal constraint to regulate payments to the original authors of books or as to who was allowed to publish a German translation of a book. Therefore, publishers were competing in translating most recent foreign works into German, which resulted in multiple German translations by different translators of the same book at the same time. To be the first one to publish a translation into German, publishers resorted to what was later called translation factories, optimized for translation speed (Bachleitner 1989). The translators working in such 'translation factories' were not specialized in the translation of one specific author. It is in fact not rare to find books from different authors translated by the same translator.

3.2.2 Method

We identified three translators who all translated books from both Shakespeare and Scott, sometimes even the same books. We also identified the English editions that were most likely to have been used by the translators. This enabled us to set up a book-level parallel English-German corpus allowing us to, again, rule out the confounding author signal.
As the constructed data set is only available in the form of PDFs from Google Books and the Internet Archive or the respective partner institutions, OCR was a necessary step for applying stylometric tools on the text corpus. To assess the quality of off-the-shelf OCR methods and to improve the OCR quality, for each book, a random set of pages was chosen for manual transcription.

3.2.3 Preparation

Following the OCR-D initiative's specifications and best practices,[10] for each book, we created a METS[11] file that contains the link to the source PDF as well as the chosen pages. The following example presents an excerpt from one of the METS files, as shown in 3.1

```
1 //...
2 <mets:fileGrp USE="IMG">
```

[10] See ocr-d spec `https://ocr-d.de/en/spec/`.
[11] See *METS. Metadata Encoding & Transmission Standard* 2022, `http://www.loc.gov/standards/mets/`.

```
3     <mets:file ID="pdf_2jMfAAAAMAAJ_28" MIMETYPE="application/pdf">
4         <mets:FLocat
5          LOCTYPE="URL"
6          xlink:href="http://books.google.com/books?id=2jMfAAAAMAAJ#page
               =28"
7          />
8     </mets:file>
9     <mets:file ID="pdf_2jMfAAAAMAAJ_103" MIMETYPE="application/pdf">
10        <mets:FLocat
11         LOCTYPE="URL"
12         xlink:href="http://books.google.com/books?id=2jMfAAAAMAAJ#page
               =103"
13         />
14    </mets:file>
15    //...
16 </mets:fileGrp>
17 //...
```

Listing 3.1: Excerpt from a METS file linking two PDF pages.

The PDFs have been downloaded from the URLs in this METS file, and the page images have been extracted from the PDF, deskewed and saved as PNG files.[12]

3.2.4 Transcription

For transcription, the standard layout analyzer of Kraken 2.0.8 (depending on the layout either with black or white column separators) has been used and the transcription was pre-filled with either the German Fraktur or the English off-the-shelf model and post-corrected manually. To ensure consistency, some characters were normalized: for example, we encountered multiple hyphenation characters such as – and = which were both transcribed by –.

3.2.5 Size

In total, the data set contains 5,354 lines with 224,745 characters. It consists of German and English books from 1815 to 1852. A detailed description of the characteristics of the data set is shown in Table 3.3.

[12]The process is implemented in the pdfs.py submodule pdfs.py:23 and it uses the command line tools imagemagick and pdfimages, see `https://github.com/millawell/ocr-data`.

3.2.6 Reproducibility and Accessibility

The data set formula has been published as a collection of PAGE files and METS files (Pletschacher and Antonacopoulos 2010). The PAGE files contain the transcriptions on line-level and the METS files serve as the container linking metadata, PDF sources and the transcriptions. There exists one METS file per item (corresponding to a Google Books or Internet Archive id) and one PAGE file per PDF page. The following excerpt of an example PAGE file shows how to encode one line of text, as shown in Listing 3.2.

```
1 //...
2 <TextLine id="textline_2">
3   <Coords points="457,124 457,1712 534,1712 534,124"/>
4   <TextEquiv>
5       <Unicode>wenn von starker Faust ein Stoß über das Schlüs-
6       </Unicode>
7   </TextEquiv>
8 </TextLine>
9 //...
```

Listing 3.2: Excerpt from a PAGE file showing the transcription of one line of text.

The `<TextLine>` contains the absolute pixel coordinates where the text is located on the preprocessed PNG image and the `<TextEquiv>` holds the transcription of the line.
As shown above, the METS files contain links to the PDFs. Additionally, the METS files contain links to the PAGE files as shown in the excerpt in Listing 3.3.

```
1  <mets:fileGrp USE="GT">
2    <mets:file ID="gt_2jMfAAAAMAAJ_28" MIMETYPE="text/xml">
3      <mets:FLocat
4       LOCTYPE="URL"
5       xlink:href="data/xml_output/2jMfAAAAMAAJ_28.page"
6      />
7    </mets:file>
8    <mets:file ID="gt_2jMfAAAAMAAJ_103" MIMETYPE="text/xml">
9      <mets:FLocat
10      LOCTYPE="URL"
11      xlink:href="data/xml_output/2jMfAAAAMAAJ_3103.page"
12     />
13   </mets:file>
14   <mets:file ID="gt_2jMfAAAAMAAJ_132" MIMETYPE="text/xml">
15     <mets:FLocat
16      LOCTYPE="URL"
17      xlink:href="data/xml_output/2jMfAAAAMAAJ_3132.page"
```

```
18        />
19     </mets:file>
20     ...
21 </mets:fileGrp>
```

Listing 3.3: Excerpt from a METS file linking to the PAGE files.

As one can see, there are links from one METS file, namely the one encoding works by Walter Scott's, Volume 2, published by the Schumann brothers in 1831 in Zwickau, identified by the Google Books id 2jMfAAAAMAAJ, to multiple pages (and PAGE files).
Finally, the METS file contains the relationship between the URLs and the PAGE files in the `<mets:structMap>` section of the file, as shown in Listing 3.4.

```
1 <mets:structMap>
2    <mets:div ID="map_001">
3        <mets:fptr FILEID="gt_2jMfAAAAMAAJ_28"/>
4        <mets:fptr FILEID="pdf_2jMfAAAAMAAJ_28"/>
5    </mets:div>
6    ...
7 </mets:structMap>
```

Listing 3.4: Excerpt from a METS file mapping the links to the PDF page and the links to the PAGE file.

In order to reuse the data set, a scholar may then obtain the original image resources from the respective institutions as PDFs, based on the links provided in the METS files. Then, the pair data set can be created by running the "make pair_output" command in the `pipelines/` directory. For each title, it extracts the PNG images from the PDF, preprocesses them, extracts, crops and saves the line images along respective files containing the text of the line.
Although the image data needs to be downloaded manually, the data set can still be compiled within minutes.

3.3 Framework for Creating, Publishing and Reusing OCR Ground-Truth Data

We have published the framework we developed for the second case study, which enables scholars to create and share their own ground truth data set formulas when they are in the same situation of not owning the copyright for the images they use. This framework offers both directions of functionality:

- Creating an XML ground truth data set from transcriptions to share it with the public (data set formula) and

- Compiling an XML ground truth data set into standard OCR ground truth data pairs to train an OCR model (built data set).[13]

As already described in the Section 3.2 there are multiple steps involved in the creation, publication and reuse of the OCR data set. In this Section, we would like to show that our work is not only relevant for scholars who want to reuse our data set but also for scholars who would like to publish a novel OCR ground truth data set in a similar copyright setting.

3.3.1 Creation and Publication

1. Corpus construction: selection of the relevant books and pages
2. Creation of the METS files[14]
3. Transcription of the pages
4. Creation of the PAGE files[15]
5. Publication of the METS and the PAGE files

3.3.2 Reuse

1. Download of the METS and PAGE files
2. Download of the PDFs as found in the METS files
3. Creation of the pair data set[16]
4. Training of the OCR models[17]

In the Section 3.2.6, the steps listed in Reuse have been described. The download of the transcriptions and the PDFs has to be done manually but for the creation of the pair data set and the training of the models, automation is provided with our framework. We would like to also automatize the download of the PDFs; this, however, remains complicated to implement. The first reason for this is a technical one: soon after starting the download, captchas appear (as early as by the 3rd image), which hinders the automatization. Another reason is the Google Books regulation itself. Page one of any Google Books PDF states explicitly:

[13]The documentation how to create a new or reproduce an existing data set can be found at README.md, `https://github.com/millawell/ocr-data/blob/master/README.md`.

[14]See mets_page_template.xml, `https://github.com/millawell/ocr-data/blob/master/data/mets_page_template.xml`.

[15]See create_xml_files.py, `https://github.com/millawell/ocr-data/blob/master/pipelines/create_xml_files.py`.

[16]See extract_pair_dataset.py, `https://github.com/millawell/ocr-data/blob/master/pipelines/extract_pair_dataset.py`.

[17]See train_ocr_model.py, `https://github.com/millawell/ocr-data/blob/master/pipelines/train_ocr_model.py`.

> Keine automatisierten Abfragen. Senden Sie keine automatisierten Abfragen irgendwelcher Art an das Google-System. Wenn Sie Recherchen über maschinelle Übersetzung, optische Zeichenerkennung oder andere Bereiche durchführen, in denen der Zugang zu Text in großen Mengen nützlich ist, wenden Sie sich bitte an uns. Wir fördern die Nutzung des öffentlich zugänglichen Materials für diese Zwecke und können Ihnen unter Umständen helfen.[18]

Finding a way to automatize download could hence not be realized in the context of this project and will have to be addressed in future work.[19] Additionally, useful templates and automation for the creation of a novel OCR ground truth data set are provided. As already described, the Kraken transcription interface was used to create the transcription. In Kraken, the final version of the transcription is stored in HTML files. We provide a script to convert the HTML transcriptions into PAGE files in order to facilitate interoperability with other OCR ground truth data sets. Finally, the pair data set can be created from the PAGE transcriptions and the images of the PDFs and the OCR model can be trained.

3.4 Relevance of the Data Set

In order to evaluate the impact that the data set has on the accuracy of OCR models, the model was trained and the performance was tested in two different settings. In the first setting, an individual model was fine-tuned for each book in the corpus using a training and an evaluation set of that book and tested the performance of the model on a held-out test set from the same book. In Table 3.3, it is shown how this data set has dramatically improved the OCR accuracy on similar documents compared to off-the-shelf OCR solutions. Especially in cases where the off-the-shelf model (baseline) shows a weak performance, the performance gained by fine-tuning is large, for example for the German translation of Woodstock by Walter Scott, the baseline accuracy was 65.91 and the fine-tuned accuracy was 94.32 which is an increase by 28.41 points.
In the second setting, the data set is split into two groups: English Antiqua, German Fraktur. There was also one German Antiqua book that was not put into any of the two groups. For the second setting, all data within a group was randomly split into train set, evaluation set and test set and an individual model was trained and tested for each group. In Table 3.2, the test performance of this setting is shown. For both groups, the fine-tuning improves the character accuracy by a large margin over

[18] When downloading any book PDF from Google Books one page is prepended to the document. On this page, the cited usage statement is presented. An English translation is: "No automatic requests. Do not send any automatic requests to the Google System. If you are researching machine translation, optical character recognition or other areas that make use of large amounts of text, please approach us directly. We support the usage of public material to that end and we may be able to help."

[19] Our progress on this topic will be documented in issue 2 of our Github repository, see `https://github.com/millawell/ocr-data/issues/2`.

the baseline accuracy. This experiment shows that overall, the fine-tuning within a group improves the performance of that group and that patterns are learned across individual books.

Document Group	**# train**	**# test**	**baseline acc.**	**fine-tuned acc.**	δ
English Antiqua	650	82	94.19	96.21	2.02
German Fraktur	3449	432	85.89	95.99	10.1

Table 3.2: Performance comparison of baseline model and fine-tuned model trained on a random splits of samples within the same group (Lassner, Coburger, et al. 2022).

A third experiment that evaluates the leave-one-out performance of the models is given in Appendix B.2.
For all three experiments, the Kraken OCR engine with a German Fraktur model and an English model was used as baselines. They were provided by the maintainers of Kraken.[20]
In the context of the research project for which this data set was created, the performance gain is especially relevant as research shows that a certain level of OCR quality is needed in order to be able to obtain meaningful results on downstream tasks. For example, Hamdi et al. 2020 show the importance of OCR quality on the performance of Named Entity Recognition as a downstream task. With additional cross training of sub-corpora We are confident that it will be possible to push the character accuracy beyond 95% on all test sets that will enable us to perform translatorship attribution analysis.
More generally, the results show that in a variety of settings, additional ground truth data will improve the OCR results. This advocates strongly for the publication of a greater range of, and especially more diverse, sets of open and reusable ground truth data for historical prints.
The data set we thus created and published is open and reproducible following the described framework. It can serve as a template for other OCR ground truth data set projects. It is therefore not only relevant because it shows why the community should create additional data sets: it also shows how to create the data sets and invites to new publications bound to bring Digital Humanities research a step forward.
The data pairs are compatible with other OCR ground truth data sets such as e. g. OCR-D (Boenig et al. 2019) or GT4HistOCR (Springmann et al. 2018) Using the established PAGE-XML standard enables interoperability and reusability of the transcriptions. Using open licenses for the source code and the data, and publishing releases at an institutional open data repository ensures representativeness and durability.

[20]See Kiessling 2019. For baselines and fine-tuning version 3.0.4 of the Kraken engine was used that can be found at `https://github.com/mittagessen/kraken/releases/tag/3.0.4`.

3.5 Summary & Conclusion

In this chapter, the opportunities for data set reuse have been discussed in the context of DH scholarship by a use case of an a novel OCR ground truth data set. In this context we presented a technical solution to publish said data set in a complex legal setting and investigated pathways for navigating the legal landscape in similar situations. Also, we showed that even small data sets on a specific language or font can have a decisive impact on the OCR quality of similar pages and documents. It was also advocated for a more consistent response from cultural heritage institution (CHI) and also for a contact person for questions regarding data set reuse to be instantiated at CHIs if non-existent.
Unfortunately, it is still not easy to publish software or data as a scholarly outcome similar to publishing more traditional scholarly works although as shown in this chapter both types of contributions can have a significant impact on the field.[21]
With regard to the OCR ground truth data set publication, one obvious limitation is that currently the process of reproducing the data set includes a manual task of downloading a list of PDFs from Google Books. This is solely due to the fact that Google Books presents captchas and explicitly discourages automatic download from Google Books. There are, however, some projects that try to get around that issue.[22]
From a mid term perspective the problem of character recognition of historical prints and handwriting with a general one-fits-all solution does not seem to be solved soon. Instead, there is a need for more and more diverse interoperable data sets that can be used to fine tune a model for a certain layout, language or script. Unfortunately there is always the risk of a new family of models emerging that will need a different kind of ground truth data. A few years back, OCR ground truth was on character level, currently, many data sets are on line level which does not make it possible to train a layout segmenter model. Therefore, nowadays also many projects push into the direction of ground truth data for whole pages. It remains to be seen what the next direction of OCR models and their corresponding ground truth data will look like. It is remarkable that there are still separate models for segmentation and recognition, resisting the general trend in machine learning toward end-to-end pipelines.

[21] The OCR ground truth data set was published as a data paper in a special issue of the Journal for Digital Humanities (Zeitschrift für digitale Geisteswissenschaften) the Standoff Converter was awarded scholarly prize. This shows that even if there are not always traditional publication schemes for data and software publications there is also hope to retrieve academic reputation through said alternative publication strategies.

[22] See `https://github.com/vaibhavk97/GoBooDo`, for example.

Title	*base model*	*# train*	*# test*	*baseline acc.*	*fine-tuned acc.*	*δ*
Tales of a Grandfather: France	Antiq.	82	11	99.8	100.0	0.2
Chronicles of the Canongate II	Antiq.	20	3	100.0	100.0	0.0
Anne of Geierstein III	Antiq.	20	3	100.0	100.0	0.0
Count Robert of Paris	Antiq.	60	8	95.54	100.0	4.46
Chronicles of the Canongate III	Antiq.	40	5	99.46	99.46	0.0
Der Alterthümler	Frak.	66	9	98.27	99.23	0.96
Quentin Durward II	Antiq.	39	5	99.17	99.17	0.0
Das gefährliche Schloß	Frak.	92	12	96.49	99.16	2.67
Walter Scott's Werke 2	Frak.	157	20	93.5	98.94	5.44
Sämmtliche dramatische Werke III	Frak.	84	11	94.5	98.85	4.35
Sämmtliche dramatische Werke IV	Frak.	125	16	92.23	98.79	6.56
Quentin Durward 15-17	Frak.	76	10	93.93	98.75	4.82
Ivanhoe	Frak.	76	10	94.58	98.45	3.87
Die schöne Mädchen von Perth	Frak.	76	10	97.19	98.31	1.12
Sämmtliche dramatische Werke VII	Frak.	77	10	92.84	98.27	5.43
J4knAAAAMAAJ	Antiq.	20	3	97.12	98.08	0.96
Woodstock 1	Frak.	52	7	95.79	98.06	2.27
Sämmtliche dramatische Werke X	Frak.	86	11	94.52	97.91	3.39
Anna von Geierstein 1-4	Frak.	88	12	93.22	97.8	4.58
Briefe über Dämonologie [..] 1-2	Frak.	71	9	94.93	97.7	2.77
Guy Mannering 3	Antiq.	20	3	96.0	97.6	1.6
Anne of Geierstein II	Antiq.	42	6	98.04	97.55	-0.49
Der Kerker von Edinburg	Frak.	76	10	91.5	97.11	5.61
Letters on Demonology [..]	Antiq.	85	11	94.73	96.7	1.97
Quentin Durward I	Antiq.	20	3	95.35	95.35	0.0
Walter Scott's Werke 7	Frak.	159	20	87.98	94.74	6.76
Woodstock	Frak.	89	12	65.91	94.32	28.41
Sämmtliche Werke in einem Bande	Frak.	1752	219	80.17	93.61	13.44
Sämmtliche dramatische Werke 6	Frak.	88	12	87.11	93.42	6.31
The Antiq.ry 2	Antiq.	61	8	90.17	92.74	2.57
The heart of Mid-Lothian I	Antiq.	19	3	91.49	92.55	1.06
Walter Scott's Romane	Frak.	183	23	71.62	91.52	19.9
Guy Mannering 2	Antiq.	36	5	88.56	90.55	1.99
Woodstock 2	Antiq.	40	6	86.78	87.6	0.82
The heart of Mid-Lothian II	Antiq.	40	6	82.72	82.72	0.0
Sämmtliche dramatische Werke XII	Frak.	73	10	68.39	79.02	10.63
Kenilworth	Frak.	78	10	69.18	78.02	8.84

Table 3.3: Performance comparison of baseline model and fine-tuned model for each document in our corpus. For almost all documents there is a large improvement over the baseline even with a very limited number of fine-tuning samples. The sum of lines and characters depicted in the table do not add up to the numbers reported in the text because during training we used an additional split of the data as an evaluation set that had the same size as the test set respectively. The best performances are highlighted for each column (Lassner, Coburger, et al. 2022).

Part II

Representing Textual Data

Chapter 4

Machine Learning Representations for Literary Text

In the first part it has been discussed what points of engagement between machine learning and digital humanities exist with a focus on the technicalities and prerequisites regarding the data sets.
The non-trivial question of defining what the text of concern is in the literary studies has been discussed and that it may depend on the research question and on the scholarly practice to decide what part of the source documents should be considered. It has been shown that TEI and the Standoff Converter enables us to extract a text version from the source document that is specific to the given research problem. These extracted text versions can be ones that are served to a human reader, for example in the form of a website of a digital critical edition. Similarly, the extracted text can be served to a machine to be transformed and represented in a specific way. These transformations are crucial for the machine to reveal 'interesting' properties of the text. This part addresses the question of how the extracted text variants can be represented and transformed with machine learning methods such that the results are meaningful in the context of traditional literary studies' methodologies.
The first chapter is focussed on the representation of textual data and the transformations of representations. The main challenge is that the methodologies in the traditional literary studies are vastly different from the ones used in machine learning. Therefore in Section 4.1, common text and word representations for ML are introduced. In Section 4.2 the basic theoretical concepts of text representations in the (computational) literary studies are established. Afterwards, methodological contact points are identified: What are the representations a human creates while reading and how can it shape a theory of machine learning reading?
One major observation is that in the literary studies there is a focus on visual representations, whereas in machine learning there can be multiple transformations of representations following each other and the intermediate representations may not be adequate for being read by a human but often the ultimate representation of such a chain of transformations is a visual representation.
Also, the difference will be discussed between machine representations that are purely algorithmic, and ones that use statistics or even machine learning. We argue

that there are chances that as soon as the transformations include machine learning, they are methodologically more similar to readings performed by a human.
It will further be pointed out what the limitations of such a machine learning model are in the context of reading machines, especially regarding the capabilities to represent the text adequately toward a specific intended reading. This goes hand in hand with the limitations in terms of unwanted biases of machine learning models. This will afterwards lead to the second chapter – Chapter 5 – of this part about text representations, where the majority of the discussed aspects are addressed by introducing the novel Word2Vec with Structure Prediction method for word representations. A method that takes into account complex structural properties of the corpus and that is able to be trained on comparatively small corpora, thereby reducing the risk of unwanted biases in the training data.

4.1 Fundamentals of Text Representations for Machine Learning

A textual representation in a very broad sense can be defined as a way of capturing features of a given text, such as a summary, a visualization or a numerical representation that can be processed by a computer. Clearly, nowadays the most widely used type of method in machine learning is a neural network. It consists of layers and each layer's (l) forward propagation can be described by a linear mapping (with weight A and bias b) and a non-linear mapping σ:

$$h_l = \sigma(A_l h_{l-1} + b_l) \tag{4.1}$$

$$h_0 = x, \tag{4.2}$$

with x being the input to the neural network.[1] This means that the input of the network has to be a tensor of real numbers and the size of the input tensor is also determined by the shape of A_1 and b_1. Therefore, in this chapter, we are seeking a textual representation that has the properties of x.
For pixel images, there is a very intuitive way to represent them as input for a neural network, namely the axes of the tensor are simply the color channels, the width and the height of the image and the float number is just the (normalized) intensity of the pixel at that position. For text, this one intuitive representation does not exist and some researchers have even tried the idea of 'printing' the text onto a pixel image to have a unified input representation for text and image, in a way reversing OCR and incorporating the OCR task into any other text-based task that one is actually interested in (Rust et al. 2022). This is clearly not the mainstream approach to text input representation for machine learning instead there are multiple common ways to represent text each having their advantages and disadvantages.
In the following, we will introduce a number of existing text representation methods with increasing complexity. The methods are all interrelated and build upon

[1] Usually, $\sigma(x) = \max(0, x)$, the ReLU activation.

one another. We will compare them by various characteristics that can help decide which representation is suitable in a given setting. An overview is given in Table 4.1, at the end of the section.

4.1.1 One-Hot Character Representation

The most naïve numerical representation of plain text would be to take the bytes as individual integer numbers. In Python this can simply be done as shown in Listing 4.1.

```
1 >>> [byte for byte in b'Hello you']
2 [72, 101, 108, 108, 111, 32, 121, 111, 117]
```

Listing 4.1: A Simple example how to convert a string to a numerical representation of bytes in Python.

With this, a matrix can be created that has as many rows as the input sequence has bytes and that has as many columns as there there are unique bytes:

$$r_{naive} = \left[\begin{array}{c|ccccccc} & \dots & 32(\sqcup), & \dots & 72(H) & ,\dots & 101(e), & \dots \\ \hline H & \dots & 0, & \dots & 1, & \dots & 0 & \dots \\ e & \dots & 0, & \dots & 0, & \dots & 1 & \dots \\ \vdots & \dots & \vdots & \dots & \vdots & \dots & \vdots & \dots \\ \vdots & \dots & \vdots & \dots & \vdots & \dots & \vdots & \dots \\ \sqcup & \dots & 1, & \dots & 0, & \dots & 0 & \dots \\ \vdots & \dots & \vdots & \dots & \vdots & \dots & \vdots & \dots \end{array}\right] \tag{4.3}$$

In this matrix, all entries are equal zero except for the position where the byte in the input sequence (rows) co-occurs with the bytes in the vocabulary (columns). Because of the sparsity property this representation is also referred to as one-hot (byte) encoding. The advantages of this representation are that any byte sequence can be represented as the procedure is not dependent on knowing words of a language and also that the sequential information of the input is preserved. A disadvantage of this representation is that it has a variable number of rows. – To input this into a neural network model, one would have to break up the rows into several chunks or pad additional rows at the bottom. Also, a small change in the data structure (a 1 shifted one position left or right) will change the representation drastically because it represents a different symbol. Another disadvantage is that the ratio between the number of rows and the number of columns is usually highly imbalanced: The 256 different symbols that can be represented with one byte stand against easily several thousand rows that will accumulate for medium to long form literary documents like a play or, of course, a novel.

4.1.2 Bag-of-Character Representation

If having a variable-length representation is problematic for the given task, the one-hot can be transformed into another popular representation, the bag of characters (BoC). This is done by taking the sum over the row axis:

$$r_{BoC} = \sum_{i}^{L} r^i_{naive} \tag{4.4}$$

with r^i, the i-th row in r. Instead of a two-axis tensor (a matrix) the representation for a document is now a one-axis tensor (a vector). This representation just tells us for example 'there are two *l*s in the document' or 'there are two *o*s in the document'. Compared to the naïve representation the bag of characters is more robust to small changes in the input sequence: Especially for longer documents, when one occurrence of a character is discarded, the count of that character over the whole document only changes slightly – and so does the bag of characters. However, this advantage comes with the cost of losing an advantage of the naïve representation: in the bag of characters representation the order of the sequence is no longer retrievable: The sequences 'cat eats fish' and 'fish eats cat' have the same bag of characters representation although having entirely different meaning.

4.1.3 Bag-of-Words Representation

As another disadvantage of the naïve one-hot character representation that was mentioned above is the balance between rows and columns. Here one could try to move some 'complexity' from the rows into the columns. For example by changing the byte encoding into to a word encoding. Instead of having a column for each byte, one could introduce a column for each word in a vocabulary. This is, of course, much more numerous – usually for English or German spanning from a few thousand to 100.000 and at the same time there are much fewer rows.[2] In this representation, locally for each word, the order of the input sequence is preserved within the columns instead of the rows. That means that even if – analogue to the bag of characters representation – one would sum over the rows, partly the order will be preserved.
Still, both the naïve word based representation (also known as one-hot word encoding) and the bag of words representation have cannot easily deal with 'new' words, may those be actual neologisms or just words that have not been considered in the vocabulary.
One way to overcome the out-of-vocabulary problem that is very popular nowadays is to find a good middle ground between bytes and words, namely subwords. The idea is that some words that occur more often should have a distinct representation but that for lower frequency words the word's representation can be compound by the subword's representations. This kind of tokenizer that decides which words

[2]For an average word length of 5 bytes, for example, one would reduce the number of rows by a factor 6, since the whitespace after a word also wouldn't have to be encoded.

should have a distinct representation, can be trained on a corpus. One very popular method is called byte-pair encoding (BPE) (Gage 1994). The training works with the following optimization algorithm: Given a sequence of bytes, collect count statistics over consecutive byte pairs (omitting word boundaries). Then, the most common byte pair will be merged into a new symbol that didn't occur in the corpus before and that will count as a byte in the next iteration. Repeat the two steps until the desired number of unique symbols is reached.[3] With this simple method, very common words are merged into one symbol hence having a distinct representation and rarer or longer words (a great example would be German compound words) still consist of multiple symbols and will therefore have a compound representation. Also out of vocabulary words can be represented with subword representations.

4.1.4 Word Embedding Representation: GloVe

Another common technique to represent text is by using so-called word embeddings. Word embeddings have several interesting properties but from one-hot representations that have been discussed so far, they differ in terms of the density of the representaion. One-hot representations are, of course, the opposite of dense because in each row is exactly one entry that is non zero. Word embeddings, in contrast, generally have no entries equal to zero. More specifically, the idea is that several syntactic and semantic properties of the words are represented in a dense embedding. In contrast to one-hot word or bag of words representations where each sample has thousands of dimensions, word embeddings only have a few dozens or a few hundred. How are these word embeddings created? There exist many different ways to create word embeddings and two of the most prominent ones are Word2Vec (Mikolov, Sutskever, et al. 2013; Bojanowski et al. 2017) and GloVe (Pennington, Socher, and Manning 2014). In this section, the GloVe method is explained but it has many similarities to Word2Vec.

For each token t in a plain text corpus (no labels needed) GloVe returns an embedding of size d. The approach is that the tokens in the text that occur in close proximity to t are markers for properties of t. As a thought experiment, if one would give a human subject ten sentences in which each time the same word is masked, would the subject be able to guess which word it is that is masked? – Many times they would. And if not the correct word, then probably a word that fits not only syntactically but also semantically and it would therefore likely be also close to the true masked word with respect to both aspects. Of course, the subject has their own representation of candidate words that comes from knowing the given language. Glove instead needs to find representations of words such that candidate words are similar to the masked word.

GloVe creates count statistics on how often a token co-occurs with all other tokens

[3] There are different variants of this algorithm, for example one where the merge is also allowed to happen over word boundaries.

across the whole training corpus.[4] The co-occurrence statistics are stored in a square matrix $Y \in \mathbb{R}^{V \times V}$ with V being the size of the vocabulary. Each row represents a center token t and each column represents a token that is in close proximity to t, a context token. After normalization, the entry of a row in Y_t yields the probability of a token to occur in the context of t. If two tokens t and t' tend to occur in similar contexts or can even be used interchangeably, then the rows Y_t and $Y_{t'}$ will be very similar. The construction of Y is already a large part of the GloVe method and the rows of Y could already be used as input representations for a neural network. However, the representation is not very memory efficient as each tokens embedding is of size V. The second step of the GloVe method is therefore to find a lower dimensional approximation of Y, $W \in \mathbb{R}^{V \times d}$. This can be done by minimizing the distance between Y and the reconstruction of W:

$$\min_{W} || \ Y - WW^T \ ||_F^2, \tag{4.5}$$

with $|| \ \cdot \ ||_F$ the Frobenius norm. The important point here is that GloVe provides d-dimensional embeddings W_t for t that still preserve the property of Y_t that syntactically or semantically similar tokens have similar embeddings.
To summarize, the GloVe method consists of

1. creating a global vocabulary of length V,
2. construct a co-occurrence matrix Y, and
3. find a d-dimensional decomposition W that minimizes the reconstruction error with d significantly smaller than V.

The decisive advantage of word embeddings over the previously discussed representations is, again, that they comprise syntactic and semantic properties of the tokens which often are very general for a language. In combination with the fact that the embeddings can be trained on unlabelled text gave rise to a particular training procedure, the pre-training pattern. For example, if one is interested in the task of named entity recognition in a specific novel from around 1830 and there exist some labelled data for one chapter of the novel. Then, the naïve approach would be to create one-hot word encodings, train a machine learning model on the chapter where labelled data is available and make predictions on the rest.
In the pre-training pattern, one could collect similar novels from that time (a 'reference corpus') to train word embeddings and then embed the labelled samples (the 'subject corpus') with these embeddings. This introduces additional general information about word use at that time to the model and therefore may improve the results (See Rusinek and Gado 2021, for example).
Nonetheless, there are also disadvantages of word embeddings. First, when used in the pre-training setting, it also often happens that unwanted information from the reference corpus is introduced into the training procedure, such as biases, and it as

[4]The name GloVe (Global Vectors) was chosen because the count statistics are collected 'globally' on the whole training corpus.

been shown that these biases may even be amplified in this pattern (Zhao et al. 2017).

4.1.5 Contextualized Representations

As the previously discussed embeddings are based solely on global statistics of word co-occurrences, the embedding of a token at a position in the corpus is not specific to the local context at that position. This is especially evident for polysemic words. One way to address this issue is to apply the pre-training pattern not only for the embeddings (the first layer of the neural network) but also apply it for larger parts of the network.

One of the early works that takes into account local context and that received widespread recognition was the ElMo method (Embeddings from Language Models, Peters et al. 2018). Instead of training individual embeddings for each word in the vocabulary, a multi-layer recurrent network is trained on input token sequences with the task of (bi-directional) next token prediction. This method does not output a fixed embedding for each word in the vocabulary but instead, to obtain the token embeddings for a given input sequence the forward pass for the whole sequence has to be computed. Due to the recurrent architecture of the network, the resulting token embeddings are dependent on the specific context of the token (in contrast to the 'global' context of previously discussed word embeddings). Using this pre-trained model as a first building block of a neural network model showed significant improvements over a variety of tasks.

Subsequently, there has been a lot of improvements in this line of research, by first, changing the architecture from a bi-directional long short-term memory neural network (BiLSTM) (Hochreiter and Schmidhuber 1997) to a transformer architecture (Vaswani et al. 2017; Radford et al. 2018; Devlin et al. 2019; Rush 2018; Phuong and Hutter 2022). Second, by scaling the pre-training phase in terms of model parameters and data set size.

One metric that is often used as a heuristic for the complexity of the model is the number of parameters. The original ElMO model has 93.6 million parameters[5], the original BERT model, which uses the transformer architecture, has 340 million parameters (Devlin et al. 2019). Interestingly, the largest collection of GloVe embeddings that are released on the official paper website also has 660 million parameters.[6][7] This is, at first, very surprising because the GloVe model was released much earlier and only consists of the embedding matrices and is therefore, intuitively much less complex than the other two models. In the BERT model, the initial embedding matrix is of size $30.522 \times 1.024 = 31.254.528$, which still makes almost ten percent of all of BERT's parameters with so much fewer types (thirty thousand compared

[5] `https://allenai.org/allennlp/software/elmo`.

[6] If we consider just one embedding matrix with a row for each word in the vocabulary and 300 columns. With a distinct embedding matrix for the context word it would even be twice that number.

[7] `https://nlp.stanford.edu/projects/glove/`, Common Crawl (840B tokens, 2.2M vocab, cased, 300d vectors, 2.03 GB download).

to more than two million for GloVe). The difference to the GloVe embeddings is, of course, the number of types in the vocabulary: with a GloVe-sized vocabulary, the embedding layer of BERT would have twice as many parameters as all transformer layers combined. Evidently, the number of parameters as a heuristic to compare models is most suitable when the model architecture is not too dissimilar.
For models with hundreds of millions of parameters, also data sets of sufficient size are needed - the particular GloVe model that was mentioned before was trained on 840 billion tokens, BERT was trained on a dataset with more than three billion tokens.
Since no labels are needed to train a GloVe or a BERT model, the previously mentioned pre-training pattern can be used: First, the model is trained on a larger corpus that does not have labels (pre-training), using a language model objective such as 'predict the masked tokens in a sentence' and afterwards, the model is trained on a data set that is specific to the research question, often with a different objective if, for example, labels are present.
This pattern has shown great results for a lot of tasks (Peters et al. 2018) however with needs for larger and larger pre-training data sets it becomes impossible to oversee and curate the data set adequately to rule out unwanted documents in the data set or any kind of biases in general (see Bender et al. 2021; Birhane and Prabhu 2021; Vries et al. 2019; Hutchinson et al. 2020).

In Table 4.1 an overview is given where the different methods can be compared across the various properties. Now, after discussing the advantages and disadvantages of the different methods, one can try to give a more informed answer to the question of which of the method might be appropriate for representing literary text in the context of a DH research question. Even if, at first glance, BERT might seem like the superior method in almost every regard, it still might be that depending on the use case, another method might is the better choice. We will revisit this question in Section 4.3 after giving more background about the methodological approaches in the computational literary studies.

4.2 Text Representations in Computation Literary Studies

In the previous section, it was discussed how text can be represented in the context of machine learning, in a sense, approaching the problem of 'text representations for machine learning reading' from the machine learning point of view. In this section, we would like to approach it from the point of view of literary studies.
Visual representations of literary phenomena have long been part of literary studies. Literary periods were represented in textbooks in the form of timelines, dramatic actions were depicted with Freytag's pyramid (Freytag 1863), and the process of interpretation was represented by the so-called hermeneutic circle (Ast 1808; Gadamer 1975) – and this has been the case since the 19th century. These approaches try to

	1-hotB	**BoC**	**BoW**	**GloVe**	**BERT**
Granularity	Byte	Document	Document	Word	Word/Sent
Dimensionality	1	100s	10000s	100s	100s
Sequence Preservation	yes	no	no	yes	yes
Local Context	no	no	no	no	yes
Input Change Sensitivity	high	high	low	high	medium
Out-of-Vocab Handled	yes	yes	no	no	yes
Row/Col Balance	no	yes	no	yes	yes
Fixed Input Size	no	yes	yes	no	no
Density	low	high	high	high	high
Number of Params	-	-	-	660mio	340mio
Pre-training Pattern	no	no	no	yes	yes

Table 4.1: Summary of the properties of the different discussed methods for text representations.

surface structures derived from the literary texts.
The idea of Distant Reading (after Moretti 2005) is thus present in these forms of text representation long before the Digital Humanities integrated them into quantitative methods in literary studies. In 1967, the semiologist Jacques Bertin developed a universal alphabet of the basic settings of visual representations (Bertin 1983)[8]. Its application in geography or sociology was the obvious first step, but its extension to literary phenomena could also be based on it in a further step. The subsequent approach by Isabelle Meirelles is a good example of embedding work from the field of visual design in other scientific contexts – including literary studies – under the auspices of information science (Meirelles 2019). In pedagogically oriented essays, Meirelles succeeds in making the basic vocabulary of visual information design legible to literary studies, i.e., in making the basic principles translatable into hermeneutic values.
Franco Moretti's work Graphs Maps Trees (Moretti 2005) sparked widespread enthusiasm for visual representations of literary texts. Besides shedding light on different patterns of representation (graphs, maps, trees), Moretti's main contribution was to see the visual representation not only as an arbitrary illustration, but as deeply coupled with text analysis. This can be seen as an intermediate step of information modeling that is important in order to use algorithmic methods on literary texts. This coupling implies that the mode of scholarship has to be made explicit – or as Johanna Drucker argues, that many visualizations "collapse the relationship between evidence and argument" (Drucker 2017).[9] Does the visualization of a text offer ways to *explore* latent structures (again, as Drucker asks, does it allow for uncertainty?) or is the visualization chosen to only support the argument that the creator was trying to make?

[8]Here, the translation is cited that was published in 1983.

[9]See also the recorded lecture by Johanna Drucker at the UCL Centre for Digital Humanities that was held as part of the the Susan Hockey lecture series, "Graphic Provocations: What do digital humanists want from visualization?", from 2016, `https://www.youtube.com/watch?v=kdey5H2Nl9w`.

The former raises the question of arbitrariness the latter the question of redundancy as it merely adds to the quantitative evidence.
Another example is given by Ramsay:

> Any reading of a text that is not a recapitulation of that text relies on a heuristic of radical transformation. The critic who endeavors to put forth a "reading" puts forth not the text, but a new text in which the data has been paraphrased, elaborated, selected, truncated, and transduced. [..] In the classroom one encounters the professor instructing his or her students to turn to page 254, and then to page 16, and finally to page 400. They are told to consider just the male characters, or just the female ones, or to pay attention to the adjectives, the rhyme scheme, images of water, or the moment in which Nora Helmer confronts her husband. The interpreter will set a novel against the background of the Jacobite Rebellion, or a play amid the historical location of the theater. He or she will view the text through the lens of Marxism, or psychoanalysis, or existentialism, or postmodernism. In every case, what is being read is not the "original" text, but a text transformed and transduced into an alternative vision, in which, as Wittgenstein put it, we "see an aspect" that further enables discussion and debate. (Ramsay 2011)

First, in this quote the term 'heuristic' in the context of a transformation is used – this connection will be elaborated on in the following section. Afterwards, reading itself is defined as a transformation of the text. This offer of a thought experiment is especially tempting in the context of digital editions as a collective effort, as discussed in Section 2.2. In this setting, every reading that yield a valuable transformation can be added, recursively to the collaborative edition as another layer.
Then, Ramsay gives a simple example of a transformation of a text that by following a set of rules that select and reorder, breaks the sequential nature of the text. Obviously, when transforming only a few pages of a book in a specific order the resulting text will be very different. The next bit of complexity that Ramsay introduces is still operationalizable, for example "pay attention to the adjectives", – if the adjectives are annotated as such in the digital edition, this can be done by an algorithm, for example by creating a new text that consists of only adjectives.
It gets more complicated toward the end of the quote as it might not be obvious how to operationalize or represent a text "through the lens of Marxism", algorithmically. This topic will, again, be picked up in the next section. The final aspect of the quote that has to be mentiooned is that Ramsay claims the transformation would "enable[..] discussion and debate". In the chapter this paragraph is quoted from, the differences between science and humanities are discussed, with an emphasis on the observation that in science, research is focussed on finding a singular answer to a problem whereas in the humanities the goal is to ask novel questions and to make sure that the discussion on the work continues. This distinction is especially important when analysing the role of visualizations in scientific works in comparison to ones from the humanitites, again supporting Drucker's argument on the evidence

of visualizations.

Often multiple high-dimensional representations are chained as intermediate representations before a final, lower-dimensional representation is retrieved and visualised for human perception.[10] This methodology is also discussed in the same work by Ramsay when he evaluates word occurrences and afterwards shows only a small subset of all the counts in a table. The important term for Ramsay is 'algorithmic', it is already in the title of the book (Toward an Algorithmic Criticism). An algorithm is a sequence of instructions. For example, given a text corpus, count the occurrences of words in each document (the so-called Bag-of-Words representation that was also introduced in the previous section) and compute the distance between each pair of documents. The intermediate representation is high dimensional (a dimension for each word in the shared vocabulary) but with the final computation of the distances, the vocabluary dimension collapses[11] and the remainder can thus be visualised – a popular choice to visualize the pair-wise distances between documents is a dendrogram.

A prominent example from the computational literary studies is stylometry using Burrows' Delta, (J. Burrows 2002). In this approach, the distances between the Bag-of-Words representations are used to distinguish between the authors of the documents in the corpus. This is especially interesting when there are documents in the corpus of unknown authorship. This method has also been used to reevaluate common assumptions on authorship of well-known books (Jannidis and Lauer 2014). The final visualisations can be interpreted by a human and therefore one could describe this type of criticism as 'algorithm-aided criticism' because the conceptual choice of which type of representation is used such that it fits the research question is done by a human being and has been done prior to using the algorithm.

Evidently, not all (intermediate) representations and not all visualisations are a good fit for every research question and the decision of how well fit these are, cannot be made by an algorithm (see also Dobson 2015).

To summarize, there exist various text representations also in non-algorithmic literary criticism. Particulary the visualization as a special type of representation is common in the humanities and there is a body of methodological discourse on this topic. The two major requirements in order to use such visualizations in the humanities is to first specify the intent: exploration or evidence for argument and also to employ the step of information modeling to make sure the transformations fit the given text corpus and the research question.

[10] A complemetary approach is suggested by Dobson 2021, namely the question of hermeneutics on the high-dimensional vector itself.

[11] Given 10 documents and 10.000 types, the intermediate representation of the corpus is a matrix with $10 \times 10.000 = 100.000$ values. The matrix can be written as $X_{i,j} \in \mathcal{R}^{10 \times 10.000}$ with i being the document index and j being the type index. Then, the pairwise distance D between documents could, for example, be computed as $D_{i,i'} = \sum_{j}^{10.000} X_{i,j} \cdot X_{i',j}$ for all $i, i' \in [0, \ldots, 10[$ so by taking the sum over the type dimension, it will collapse to a single value in the resulting distance matrix.

4.3 Machine Learning Reading

In this section, it will be discussed in which ways the concept of representation can be interchanged or harmonized between the two disciplines, and what methodological implications it has when machine learning representations take the place of the representations previously used in traditional literary studies. In the previous section, it was shown that algorithmic transformations have also been used in traditional literary studies, and that they are part of the methodological discourse. In addition, the use of algorithmic transformations has opened up the possibility of new methods in literary studies, as has been demonstrated extensively (Baillot 2018; Moretti 2005). The question that goes beyond all this, however, is: do methods based on ML representations give us the possibility of imitating or adopting even more (and possibly more complex) components of the existing methodology of literary studies? We do not claim that by applying machine learning methods, literary scholars could suddenly be replaced by statistical models, but the clear boundaries between the reading machine and the reading human shift a bit when statistical representations are used instead of rule-based ones. For example, instead of a bag-of-words representation, as discussed earlier, word embeddings could also be used. Word embeddings (Mikolov, Chen, et al. 2013; Pennington, Socher, and Manning 2014) refer to representations in which each word in the corpus is associated with a vector. Here, the representations of the words in the vector space are arranged in such a way that words that are syntactically or semantically similar to each other are also close to each other in the represented space. These word embeddings do not necessarily have to be trained on a corpus that is to be analyzed but can also be trained on another corpus and the found word representations can be used for representing the corpus that is to be analyzed (we call the corpus to be analyzed the subject corpus and call the corpus on which the word embeddings were trained the reference corpus) and thus the word embeddings contain certain syntactic and semantic information of the reference corpus.

Word embeddings can, for example, be used to identify variants of named entities (Rusinek and Gado 2021) or be used to measure abstraction of words in fiction (Heuser 2020). For example, with word embeddings trained on German plays from the late seventeen hundreds, words with certain syntactic properties could also be identified, such as imperative forms in a play that the embeddings were not trained on. A visualization of the results based on these representations would then no longer be purely rule-based, nor would they be based only on the subject corpus, but would be a selective aggregation of the information extracted from the reference corpus. Distinctively, one could speak of an evolution from ‘rule-based’ (algorithm), to ‘based only on the subject corpus’ (classical statistics), to ‘based on reference corpus’ (machine learning).

One could see this interaction between the large reference corpus and the subject corpus under investigation analogue to the interaction between readers of a text

(subject corpus) and their personal reading history, the 'reading horizon'[12] (reference corpus). When in the previous section, the question was left unanswered, how certain readings in the sense of radical transformations may be operationalizable, such as reading "through the lens of Marxism", an answer would be to first create textual representations, such as word embeddings on a reference corpus that comprises Marxist literature and use the representations for the subject corpus to create a specific reading.

The conceptual, or information modeling step is still needed in this approach, only other considerations have to be made, e.g. 'How to construct the reference corpus?', "What properties should be included in the representation, - should one use only lemmas of the reference corpus, for example?". So, as before, much of the choice of methods and parameters remains in the hands of the researchers. Furthermore, despite intensive research by the machine learning community, the problem of representing longer text sequences still exists today. Even though embeddings have not been limited to single words like word embeddings any more, instead many models (most notably Large Language Models that were mentioned before, Devlin et al. 2019) using the widely used transformer architecture (Vaswani et al. 2017) can handle sequences no more than a few hundred words long. However, in the machine learning model's reading horizon proposed above, there is also a danger for many applications: if the reference corpus is poorly documented or is so large that it cannot be documented cleanly at all, it is equally difficult to discern which biases have been adopted by the trained model. This criticism was summed up by Bender et al. 2021. Dangers of perpetuating or reinforcing these biases exist especially for users of modern NLP applications, where it is often not even public what they were trained on. Researchers should be aware of these dangers in their work. However, there are also opportunities to gain insights: in the context of simple word embeddings, it has already been shown that analyzing biases in parts of the reference corpora used can shed light on important insights about social stereotypes of certain times (Garg et al. 2018). Ted Underwood also argues along these lines in his recent paper (Underwood 2022), formulating that "The immediate value of these models [i.e., Language Models] is often not to mimic individual language understanding, but to represent specific cultural practices (like styles or expository templates) so they can be studied and creatively remixed." And Underwood goes on to say that "[w]hen research is organized by this sort of comparative purpose, the biases in data are not usually a reason to refrain from modeling – but a reason to create more corpora and train models that reflect a wider range of biases." A typical approach would be to train a general language model on a large corpus of the present time (pre-training) and then to focus it on other historical contexts (fine tuning). The problem with this is that it is not well documented (or documentable at the moment) how much of the contemporary corpus is still present in the model after fine tuning that does not

[12]The term 'reading horizon' is used although this is not a common term in English. In German, the term 'Lesehorizont' is established which refers to all the works that are known to the reader, or, to stay in within the metaphor – everything from the literary landscape that is visible to the reader. This is the corpus that the reader has at their disposal to decrypt literary references or intertextualities or to apply comparative methods.

correspond to the historical context of use. At the same time, there are probably few historical contexts today for which sufficiently large text corpora exist to allow training without pre-training of models that are as data hungry as the transformer models.

4.4 Summary & Conclusion

The goal of this chapter was to identify how to use machine learning methods in the computational literary studies in a meaningful way from a methodological perspective. On the precondition that both disciplines have to align their methodologies, first the fundamentals of text representations in machine learning were introduced and an overview over the discourse of the methodology of text representations, transformations and visualizations from the computational literary studies was given. The two main observations were that typically in the CLS, one has to explicitly differentiate between explorative and confirmatory experiments. Therefore, the literary critic may use ML methods either for exploratory or confirmatory analyses. Optimally, scholarship is supported by both, exploratory and confirmatory experiments as it will be shown in the case studies in the final part of this thesis.
Secondly, only a fraction of the types of transformations that are discussed by Ramsay as being possible methodological readings of literature can be easily operationalized with purely algorithmic transformations. Therefore, the methodological contact point between 'training a textual representation on a reference corpus' and a specific 'reading horizon' was identified showing how machine learning in particular enables the CLS methodology to advance, moving forward.
Of course, ML methods have not replaced literary critics and they presumably will not in the forseable future. Besides the fact that the procedures of an ML method is very different to what a literary critic does, in order to take on meaningful work in computational literary studies with machine learning the important step of information modeling and chosing the right methodology remains in the hand of the (human) critic.
In addition, the usefulness of the idea of a reading machine to its own end is somewhat limited because, actually similar to a machine that writes, literature and all actors that are traditionally involved (author, translator, editor, critic, reader, to name a few) usually form a feedback loop and interact with the socio-cultural setting it is taking place in. Therefore, replacing actors in this setting might make the whole endeavour more and more pointless, until the absurdity of a writing machine hard-wired to a reading machine as an autonomous system, reading and writing on their own, ad infinity.
Another limitation is that there exist many types of representations and these have an influence on what is actually represented as reading horizon. This approach also bears the risk of unwanted reading horizons as with growing sizes of the reference corpora, they also become harder to curate.

Also, engaging in this type of information modeling requires the literary scholar to make decisions about machine learning and such is yet not widely taught in literature classes at universities.

Chapter 5

Word Representations for Structured Corpora

In the previous chapter, it was argued that word representations that are trained on a specific set of sources, methodologically, can be seen as corresponding to the reading horizon of a literary critic. So far, the approach that was discussed, was training word embeddings on a specific reference corpus, for example, all of the works of one author to create representations characteristic for this author, and for a different experiment one might use all of the works of a different author to train representations. The issue with this approach is that, of course, many of the words that a representation is created from when training for author one have a very similar meaning in the sub-corpus defined by author two. The majority in deviation of meaning can be assumed to be mostly concerning a small subset of words. It is therefore inefficient to train word representations for each sub-corpus (for each author) individually. Very similar to the argument that was given in the context of textual variants, where it was argued that not only specific variants could be an interesting subject of research but actually the comparison between different variants. Here, it might be as interesting from a digital humanities point of view to have adequate representations for a specific author as it might be to investigate the similarities between different authors, or in other words, the structure of the representations of sub-corpora.

We propose an application of the novel methods to the field of digital humanities, and develop an example more specifically related to computational literary studies. In the renewal of literary studies brought by the development and implementation of computational methods, questions of authorship attribution and genre attribution are key to formulating a structured critique of the classical design of literary history, and of cultural heritage approaches at large. In particular the investigation of historical person networks, knowledge distribution and intellectual circles has shown to benefit significantly from computational methods (Baillot 2018; Moretti 2005). Hence, the methods presented in this chapter and its capability to reveal connections between sub-corpora (such as authors' works), can be applied with success to these types of research questions. Here, the use of quantitative and statistical models can lead to new, hitherto unfathomed insights. A corpus-based statistical approach to

literature also entails a form of emancipation from literary history in that it makes it possible to shift perspectives, e.g., to reconsider established author-based or genre-based approaches.

There are many situations, not only in the context of digital humanities, where a given target corpus is considered to have some *structure*. For example, when analyzing news articles, one can expect that articles published in 2000 and 2001 are more similar to each other than the ones from 2000 and 2010. When analyzing scientific articles, uses of technical terms are expected to be similar in articles on similar fields of science. This implies that the structure of a corpus can be a useful side information for obtaining better word representation.[1]

We argue that apart from diachronic word embeddings there is a need to train dynamic word embeddings that not only capture temporal shifts in language but for instance also semantic shifts between domains or regional differences. It is therefore important that those embeddings can be trained on small datasets.

We therefore propose in Lassner, Brandl, et al. 2023 two methods. The first method is called *Word2Vec with Structure Constraint (W2VConstr)*, where domain-specific embeddings are learned under regularization with any kind of structure. This method performs well when a respective graph structure is given a priori. For more general cases where no structure information is given, we propose a second method, called *Word2Vec with Structure Prediction (W2VPred)*, where domain-specific embeddings and subcorpora structure are learned at the same time. W2VPred simultaneously solves three central problems that arise with word embedding representations:

1. Words in the sub-corpora are embedded in the same vector space, and are therefore directly comparable without post-alignment.

2. The different representations are trained simultaneously on the whole corpus as well as on the sub-corpora, which makes embeddings for both general and domain-specific words robust, due to the information exchange between sub-corpora.

3. The estimated graph structure can be used for confirmatory evaluation when a reasonable prior structure is given. W2VPred together with W2VConstr identifies the cases where the given structure is not ideal, and suggests a refined structure which leads to an improved embedding performance, we call this method *Word2Vec with Denoised Structure Constraint*. When no structure is given, W2VPred provides insights on the structure of sub-corpora, e.g., similarity between authors or scientific domains.

The presented methods rely on static word embeddings as opposed to currently often used contextualized word embeddings. As we learn one representation per slice such as year or author, thus considering a much broader context than contextualized embeddings, we are able to find a meaningful structure between corresponding slices.

[1] In the case of image categorization, the usefulness of supplementary taxonomies during training has been discussed by Binder, K.-R. Müller, and Kawanabe 2012.

Another main advantage comes from the fact that the proposed methods do not require any pre-training and can be run on a single GPU.
The methods are tested on 4 different datasets with different structures (sequences, trees and general graphs), domains (news, wikipedia, high literature) and languages (English and German). We show on numerous established evaluation methods that W2VConstr and W2VPred significantly outperform baseline methods with regard to general as well as domain-specific embedding quality. We also show that W2VPred is able to predict the structure of a given corpus, outperforming all baselines. Additionally, we show robust heuristics to select hyperparameters based on proxy measurements in a setting where the true structure is not known. Finally, we show how W2VPred can be used in an explorative setting to raise novel research questions in the field of digital humanities.

5.1 Related Work on Corpus Structure and Word Embedding Dynamics

Various approaches to track, detect and quantify semantic shifts in text over time have been proposed (Sugiyama, Krauledat, and K.-R. Müller 2007; Kim et al. 2014; Kulkarni et al. 2015; Hamilton, Leskovec, and Jurafsky 2016; Zhang et al. 2016; Marjanen et al. 2019).
This research is driven by the hypothesis that semantic shifts occur, e.g., over time (Bleich, Nisar, and Abdelhamid 2016) and viewpoints (Azarbonyad et al. 2017), in political debates (Reese and Lewis 2009) or caused by cultural developments (Lansdall-Welfare et al. 2017).
Typically, methods first train individual static embeddings for different timestamps, and then align them afterwards (e.g., Kulkarni et al. 2015; Hamilton, Leskovec, and Jurafsky 2016; Kutuzov et al. 2018; Devlin et al. 2019; Jawahar and Seddah 2019; Hofmann, Pierrehumbert, and Schütze 2020) which is also discussed in a comprehensive survey by (Tahmasebi, Borin, and Jatowt 2018). Other approaches, which deal with more general structure (Azarbonyad et al. 2017; Gonen et al. 2020) and more general applications (Zeng et al. 2017; Shoemark et al. 2019), also rely on post-alignment of static word embeddings (Grave, Joulin, and Berthet 2019). With the rise of large language models that use contextualized embeddings, a part of the research question has shifted towards detecting language change in contextualized word embeddings (e.g., Jawahar and Seddah 2019; Hofmann, Pierrehumbert, and Schütze 2020).
Recent methods directly learn dynamic word embeddings in a common vector space without post-alignment: Bamler and Mandt 2017 proposed a Bayesian probabilistic model that generalizes the skip-gram model (Mikolov, Sutskever, et al. 2013) to learn dynamic word embeddings that evolve over time. Rudolph and D. Blei 2018 analyzed dynamic changes in word embeddings based on exponential family embeddings, a probabilistic framework that generalizes the concept of word embeddings to other types of data (Rudolph, F. Ruiz, et al. 2016). Yao et al. 2018 proposed Dynamic

Word2Vec (DW2V) to learn individual word embeddings for each year of the New York Times dataset (1990-2016) while simultaneously aligning the embeddings in the same vector space. Specifically, they solve the following problem for each timepoint $t = 1, \ldots, T$ sequentially:

$$\min_{U_t} L_{\mathrm{F}} + \tau L_{\mathrm{R}} + \lambda L_{\mathrm{D}}, \quad \text{where} \tag{5.1}$$

$$L_{\mathrm{F}} = \left\| Y_t - U_t U_t^\top \right\|_F^2, L_{\mathrm{R}} = \|U_t\|_F^2,$$

$$L_{\mathrm{D}} = \|U_{t-1} - U_t\|_F^2 + \|U_t - U_{t+1}\|_F^2 \tag{5.2}$$

represent the losses for data fidelity, regularization, and diachronic constraint, respectively. $U_t \in \mathbb{R}^{V \times d}$ is the matrix consisting of d-dimensional embeddings for V words in the vocabulary, and $Y_t \in \mathbb{R}^{V \times V}$ represents the positive pointwise mutual information (PPMI) matrix (Levy and Goldberg 2014). The diachronic constraint L_{D} encourages alignment of the word embeddings with the parameter λ controlling how much the embeddings are allowed to be dynamic ($\lambda = 0$: no alignment and $\lambda \to \infty$: static embeddings).

5.2 Methods

By generalizing DW2V, we propose two methods, one for the case where sub-corpora structure is given as prior knowledge, and the other for the case where no structure is given a priori. As previously said, this makes sense in the context of the methodologies of digital humanities research, employing both confirmatory (structure is given) and explorative (no structure given) methods.
We also argue that combining both methods can improve the performance in cases where some prior information is available but not necessarily reliable.

5.2.1 Word2Vec with Structure Constraint

We reformulate the diachronic term in Eq. 5.1 as

$$L_{\mathrm{D}} = \sum_{t'=1}^{T} W_{t,t'}^{\mathrm{diac}} \|U_t - U_{t'}\|_F^2$$

$$\text{with } W_{t,t'}^{\mathrm{diac}} = \mathbb{1}(\{|t - t'| = 1\}), \tag{5.3}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. This allows us to generalize DW2V for different neighborhood structures: Instead of the chronological sequence (5.3), we assume $W \in \mathbb{R}^{T \times T}$ to be an arbitrary affinity matrix representing the underlying semantic structure, given as prior knowledge.
Let $D \in \mathbb{R}^{T \times T}$ be the pairwise distance matrix between embeddings such that

$$D_{t,t'} = \|U_t - U_{t'}\|_F^2, \tag{5.4}$$

and we impose regularization on the distance, instead of the norm of each embeddings. This yields the following optimization problem:

$$\min_{U_t} L_{\mathrm{F}} + \tau L_{\mathrm{RD}} + \lambda L_{\mathrm{S}}, \quad \text{where} \tag{5.5}$$

$$L_{\mathrm{F}} = \left\| Y_t - U_t U_t^\top \right\|_F^2, L_{\mathrm{RD}} = \|D\|_F,$$

$$L_{\mathrm{S}} = \sum_{t'=1}^{T} W_{t,t'} D_{t,t'}. \tag{5.6}$$

We call this generalization of DW2V *Word2Vec with Structure Constraint* (W2VConstr).

5.2.2 Word2Vec with Structure Prediction

When no structure information is given, we need to estimate the similarity matrix W from the data. We define W based on the similarity between embeddings. Specifically, we initialize (each entry of) the embeddings $\{U_t\}_{t=1}^T$ by independent uniform distribution in $[0, 1)$. Then, in each iteration, we compute the distance matrix D by Eq.(5.4), and set $\widetilde{W}$ to its (entry-wise) inverse, i.e.,

$$\widetilde{W}_{t,t'} \leftarrow \begin{cases} D_{t,t'}^{-1} & \text{for } t \neq t', \\ 0 & \text{for } t = t'. \end{cases} \tag{5.7}$$

and normalize it according to the corresponding column and row:

$$W_{t,t'} \leftarrow \frac{\widetilde{W}_{t,t'}}{\sum_{t''} \widetilde{W}_{t,t''} + \sum_{t''} \widetilde{W}_{t'',t'}}. \tag{5.8}$$

The structure loss (5.6) with the similarity matrix W updated by Eqs. 5.7 and 5.8 constrains the distances between embeddings according to the similarity structure that is at the same time estimated from the distances between embeddings. We call this variant *Word2Vec with Structure Prediction* (W2VPred). Effectively, W serves as a weighting factor that strengthens connections between close embeddings.

5.2.3 Word2Vec with Denoised Structure Constraint

We propose a third method that combines W2VConstr and W2VPred for the scenario where W2VConstr results in poor word embeddings because the a-priori structure is not optimal. In this case, we suggest to apply W2VPred and consider the resulting structure as an input for W2VConstr. This procedure needs prior knowledge of the dataset and a human-in-the-loop to interpret the predicted structure by W2VPred in order to add or remove specific edges in the *new* ground truth structure. In the experiment section, we will condense the predicted structure by W2VPred into a sparse, denoised ground truth structure that is meaningful. We call this method Word2Vec with Denoised Structure Constraint (W2VDen).

5.2.4 Optimization

We solve the problem (5.5) iteratively for each embedding U_t, given the other embedings $\{U_{t'}\}_{t' \neq t}$ are fixed.

We define one epoch as complete when $\{U_t\}$ has been updated for all t. We applied gradient descent with Adam (Kingma and Ba 2014) with default values for the exponential decay rates given in the original paper and a learning rate of 0.1. The gradients at timestep i are given by

$$g_t^i = \nabla(L_{F(t)} + \tau L_{RD} + \lambda L_{S(t)})_{U^i} \tag{5.9}$$

The learning rate has been reduced after 100 epochs to 0.05 and after 500 epochs to 0.01 with a total number of 1000 epochs. Both models have been implemented in PyTorch. W2VPred updates W by Eqs. 5.7 and 5.8 after every iteration.

5.3 Establishing Novel Methods on Benchmark Data

In this chapter, novel methods for word representations subject to corpus structure have been introduced. Before they can confidently be used in the context of DH experiments where the outcome is unknown, the validity of the methods have to be shown on established data sets and their performance has to be compared to existing methods. It is, as discussed in Section 5.1 uncommon yet for existing methods to yield both, word representations and a predicted affinity matrix for sub-corpora. The performance of the presented methods will therefore be compared against different baseline methods in each task individually.

There are three tasks that we evaluate the method against. First, we evaluate general embedding performance which corresponds to general characteristics of the words. – As discussed before, we assume that most of the words have similar characteristics across sub corpora and only few words change their semantics or the way they are syntactically used for different sub-corpora. First, general analogies (Mikolov, Sutskever, et al. 2013) are used to evaluate the methods. In a second step, the methods are evaluated on a set of common word similarity tasks (Faruqui and Dyer 2014). Finally, the embeddings are analyzed with QVEC, a measure of component-wise correlation (Tsvetkov et al. 2015).

The second task that we evaluate the methods against is the domain-specific embedding performance. For this, an existing set of analogy tests is used that is subject to temporal change (Yao et al. 2018; Szymanski 2017).

Finally, the methods are evaluated on structure prediction performance. For this, structure labels are retrieved from the meta data of the data sets and are compared against the predicted structure.

Category	#Articles
Natural Sciences	8536
Chemistry	19164
Computer Science	11201
Biology	10988
Engineering & Technology	20091
Civil Engineering	17797
Electrical & Electronic Engineering	6809
Mechanical Engineering	4978
Social Sciences	17347
Business & Economics	14747
Law	13265
Psychology	5788
Humanities	15066
Literature & Languages	24800
History & Archaeology	16453
Religion & Philosophy & Ethics	19356

Table 5.1: Categories and the number of articles in the WikiFoS dataset. One cluster contains 4 categories (rows): the top one is the main category and the following 3 are subcategories. Fields joined by & originate from 2 separate categories in Wikipedia[4] but were joined, according to the OECD's definition.[3]

In the following subsections, we will first describe the data, preprocessing and then the results. Further details on implementation and hyperparameters can be found in Appendix C.1.

5.3.1 Datasets

We evaluated the methods on the following three benchmark datasets. The details on preprocessing are given in Appendix C.2.

New York Times (NYT): The New York Times dataset[2] (NYT) contains headlines, lead texts and paragraphs of English news articles published online and offline between January 1990 and June 2016 with a total of 100,945 documents. We grouped the dataset by years with 1990-1998 as the train set and 1999-2016 as the test set.

Wikipedia Field of Science and Technology (WikiFoS): We selected categories of the OECD's list of Fields of Science and Technology[3] and downloaded the corresponding articles from the English Wikipedia. The resulting dataset Wikipedia Field of Science and technology (WikiFoS) contains four clusters, each of which consists of one main category and three subcategories, with 226,386 unique articles in total (see Table 5.1). The articles belonging to multiple categories[4] were randomly assigned to a single category in order to avoid similarity because of overlapping

[2]See `https://sites.google.com/site/zijunyaorutgers/`.
[3]`http://www.oecd.org/science/inno/38235147.pdf`
[4]`https://en.wikipedia.org/wiki/Wikipedia:Contents/Categories`

Category	#Articles
Logic	3394
Concepts in Logic	1455
History of Logic	76
Aesthetics	7349
Philosophers of Art	30
Literary Criticism	3826
Ethics	5842
Moral Philosophers	170
Social Philosophy	3816
Epistemology	3218
Epistemologists	372
Cognition	8504
Metaphysics	1779
Ontology	796
Philosophy of Mind	976

Table 5.2: Categories and the number of articles in the WikiPhil dataset. One cluster contains 3 categories: the top one is the main category and the following are subcategories in Wikipedia.

texts instead of structural similarity. In each category, we randomly chose 1/3 of the articles for the train set, and the remaining 2/3 were used as the test set. In the case of this data set and also the WikiPhil data set, no temporal order is used.

Wikipedia Philosophy (WikiPhil): Based on Wikipedia's definition of categories in *philosophy*, we selected 5 main categories and their 2 largest subcategories each (see Table 5.2). Categories and subcategories are based on the definition given by Wikipedia. We downloaded 41,603 unique articles in total from the English Wikipedia. Similarly to WikiFoS, the articles belonging to multiple categories were randomly assigned to a single category, and the articles in each category were divided into a train set (1/3) and a test set (2/3).

5.3.2 Baseline Methods

For the general embedding quality, four different baseline methods are compared against: GloVe, that was discussed in Section 4.1, DW2V that was discussed in Section 5.1 and the two variants of the Word2Vec method (Mikolov, Sutskever, et al. 2013), CBOW and Skip-Gram are used.
For domain specific embedding quality, the same baseline methods are used however only DW2V is expected to have comparable results as the others are not modeling the corpus structure and are rather reported for completeness.
For evaluating the structure prediction performance, Burrows' Delta is added as another baseline. This method does not yield word embeddings similar to the other methods but it is a well-known method in the digital humanities and it can be used for sub-corpus structure prediction.

Data	Method	general analogy tests		
		n=1	n=5	n=10
NYT	GloVe	9.40	26.41	33.58
	Skip-Gram	3.62	16.20	25.61
	CBOW	5.58	19.92	27.60
	DW2V	11.27	32.88	42.97
	W2VConstr (our)	10.90	33.01	43.12
	W2VPred (our)	10.28	31.66	41.88
WikiFoS	GloVe	6.33	23.74	32.58
	Skip-Gram	3.54	12.09	15.77
	CBOW	4.25	17.47	26.21
	W2VConstr (our)	11.91	45.96	56.88
	W2VPred (our)	11.82	45.73	56.40
	W2VDen (our)	11.61	46.50*	57.08*
WikiPhil	GloVe	2.59	17.45	24.19
	Skip-Gram	2.76	10.18	17.48
	CBOW	3.11	6.61	9.47
	W2VConstr (our)	0.42	10.37	15.02
	W2VPred (our)	4.37	31.99	41.75
	W2VDen (our)	5.96*	36.21*	46.15*

Table 5.3: General analogy test performance for our methods, W2VConstr and W2VPred, and baseline methods, GloVe, Skip-Gram, CBOW and DW2V averaged across ten runs with different random seeds. The best method and the methods that are not significantly outperformed by the best is marked with a gray background, according to the Wilcoxon signed rank test for $\alpha = 0.05$. W2VDen is compared against the best method from the same data set and if it is significantly better, it is marked with a star (*).

5.3.3 General Embedding Performance

In the first experiment, we compare the quality of the word embeddings trained by W2VConstr and W2VPred with the embeddings trained by baseline methods, GloVe, Skip-Gram, CBOW and DW2V. For GloVe, Skip-Gram and CBOW, we computed one set of embeddings on the entire dataset. For DW2V, W2VConstr and W2VPred, domain-specific embeddings $\{U_t\}$ were averaged over all domains. We use the same vocabulary for all methods. For W2VConstr, we set the affinity matrix W as shown in the upper row of Figure 5.1, based on the a priori known structure, i.e., diachronic structure for NYT, and the category structure in Tables 5.1 & 5.2 for WikiFoS and WikiPhil. The lower row of Figure 5.1 shows the learned structure by W2VPred.

Specifically, we set the ground-truth affinity $W^*_{t,t'}$ as follows: for NYT, $W^*_{t,t'} = 1$ if $|t - t'| = 1$, and $W^*_{t,t'} = 0$ otherwise; for WikiFoS and WikiPhil, $W^*_{t,t'} = 1$ if t is the parent category of t' or vice versa, $W^*_{t,t'} = 0.5$ if t and t' are under the same parent category, and $W^*_{t,t'} = 0$ otherwise (see Tables 5.1 and 5.2 for the category structure of WikiFoS and WikiPhil, respectively, and the top row of Figure 5.1 for the visualization of the ground-truth affinity matrices).

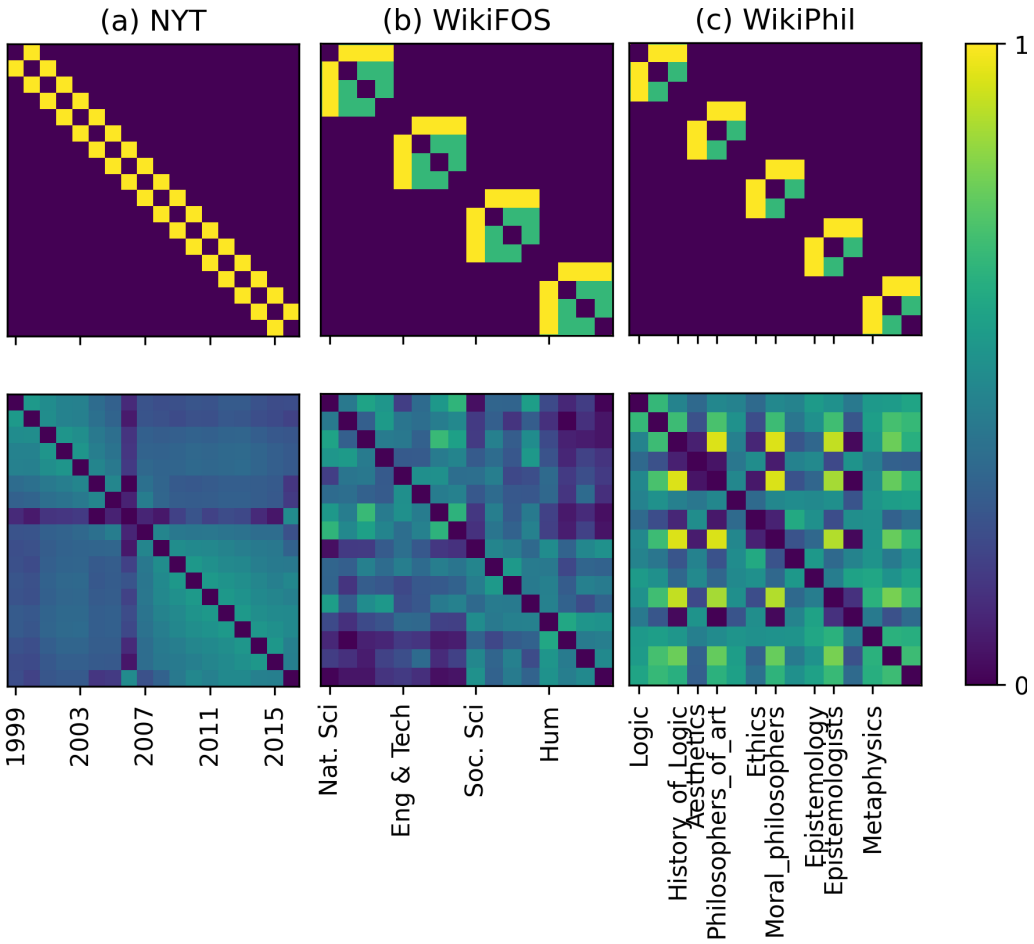

Figure 5.1: Prior affinity matrix W used for W2VConstr (upper), and the estimated affinity matrix by W2VPred (lower) where the number indicates how close slices are (1: identical, 0: very distant). The estimated affinity for NYT implies the year 2006 is an outlier. We checked the corresponding articles and found that many paragraphs and tokens are missing in that year. Note that the diagonal entries do not contribute to the loss for all methods.

We evaluate the embeddings on general analogies (Mikolov, Sutskever, et al. 2013) to capture the general meaning of a word. Table 5.3 shows the corresponding accuracies averaged across 10 runs with different random seeds.

For NYT, W2VConstr performs similarly to DW2V, which has essentially the same constraint term—L_S in Eq.(5.6) for W2VConstr is the same as L_D in Eq.(5.2) for DW2V up to scaling when W is set to the prior affinity matrix for NYT—and significantly outperforms the other baselines. W2VPred performs slightly worse then the best methods. For WikiFoS, W2VConstr and W2VPred outperform all baselines by a large margin. In WikiPhil, W2VConstr performs poorly (worse than GloVe), while W2VPred outperforms all other methods by a large margin. Standard deviation across the 10 runs are less than one for NYT (all methods and all n), slightly higher for WikiFoS and highest for WikiPhil W2VPred and W2VConstr (0.28-3.17). These different behaviors can be explained by comparing the estimated (lower row) and the a priori given (upper row) affinity matrices shown in Figure 5.1. In NYT, the estimated affinity decays smoothly as the time difference between two slices increases.

This implies that the a priori given diachronic structure is good enough to enhance the word embedding quality (by W2VConstr and DW2V), and estimating the affinity matrix (by W2VPred) slightly degrades the performance due to the increased number of unknown parameters to be estimated. In WikiFoS, although the estimated affinity matrix shows somewhat similar structure to the given one a priori, it is not as smooth as the one in NYT and we can recognize two instead of four clusters in the estimated affinity matrix consisting of the first two main categories (*Natural Sciences* and *Engineering & Technology*), and the last two (*Social Sciences* and *Humanities*), which we find reasonable according to Table 5.1. In summary, W2VConstr and W2VPred outperform baseline methods when a suitable prior structure is given. Results on the WikiPhil dataset show a different tendency: the estimated affinity by W2VPred is very different from the prior structure, which implies that the corpus structure defined by Wikipedia is not suitable for learning word embeddings. As a result, W2VConstr performs even poorer than GloVe. Overall, Table 5.3 shows that the proposed W2VPred robustly performs well on all datasets. In Section 5.3.5, we will further improve the performance by *denoising* the estimated structure by W2VPred for the case where a prior structure is not given or unreliable.
We further evaluate word embeddings on various word similarity tasks where human-annotated similarity between words is compared with the cosine similarity in the embedding space, as proposed in Faruqui and Dyer 2014. Table 5.7 shows the correlation coefficients between the human-annotated similarity and the embedding cosine similarity, where, again, the best method and the runner-ups (if not significantly outperformed) are highlighted.[5] We observe that W2VPred outperforms the other methods in 7 out of 12 datasets for NYT, and W2VConstr in 8 out of 12 for WikiFoS. For WikiPhil, since we already know that W2VConstr with the given affinity matrix does not improve the embedding performance, we instead evaluated W2VDen, which outperforms 9 out of 12 datasets in WikiPhil. In addition, W2VPred gives comparable performance to the best method over all experiments.
We also apply QVEC which measures component-wise correlation between distributed word embeddings, like we use them throughout this section, and linguistic word vectors based on WordNet Fellbaum 1998. High correlation values indicate high saliency of linguistic properties and thus serve as an intrinsic evaluation method that has been shown to highly correlate with downstream task performance Tsvetkov et al. 2015. Results are shown in Table 5.8, where we observe that W2VConstr (as well as W2VDen for WikiPhil) outperforms all baseline methods, except CBOW in NYT, on all datasets, and W2VPred performs comparably with the best method.

5.3.4 Domain-Specific Embedding Performance

Yao et al. 2018 and Szymanski 2017 introduced temporal analogy tests that allow us to assess the quality of word embeddings with respect to their temporal information.

[5] We removed the dataset *VERB-143* since we are using lemmatized tokens and therefore catch only a very small part of this corpus. We acknowledge that the human annotated similarity is not domain-specific and therefore not optimal for evaluating the domain-specific embeddings. However, we expect that this experiment provides another aspect of the embedding quality.

	n=1	n=5	n=10
GloVe	7.72	14.39	17.87
Skip-Gram	10.49	19.89	24.78
CBOW	6.35	11.36	14.59
DW2V	39.47	61.94	67.35
W2VConstr (our)	38.23	57.73	64.54
W2VPred (our)	41.87	64.60	69.67

Table 5.4: Accuracies for temporal analogies (NYT).

Unfortunately, domain-specific tests are only available for the NYT dataset. Table 5.4 shows temporal analogy test accuracies on the NYT dataset. As expected, GloVe, Skip-Gram and CBOW perform poorly. We assume this is because the individual slices are too small to train reliable embeddings. The embeddings trained with DW2V and W2VConstr are learned collaboratively between slices due to the diachronic and structure terms and significantly improve the performance. Notably, W2VPred further improves the performance by learning a more suitable structure from the data. Indeed, the learned affinity matrix by W2VPred (see Figure 5.1a) suggests that not the diachronic strcuture used by DW2V but a smoother structure is optimal.

Nat. Sci	Eng&Tech	Soc. Sci	Hum	GloVe	Skip-Gram
generator	generator	powerful	powerful	control	Power
PV	inverter	control	control	supply	inverter
thermoelectric	alternator	wield	counterbalance	capacity	mover
inverter	converter	drive	drive	system	electricity
converter	electric	generator	supreme	internal	thermoelectric

Table 5.5: Five nearest neighbors to the word "power" in the domain-specific embedding space, learned by W2VPred, of four main categories of WikiFoS (left four columns), and in the general embedding space learned by GloVe and Skip-Gram on the entire dataset (right-most columns, respectively).

Since no domain-specific analogy test is available for WikiFoS and WikiPhil, we qualitatively analyzed the domain-specific embeddings by checking nearest neighboring words. Table 5.5 shows the 5 nearest neighbors of the word "power" in the embedded spaces for the 4 main categories of WikiFoS trained by W2VPred and GloVe and Skip-Gram. We averaged the embeddings obtained by W2VPred over the subcategories in each main category. The distance between words are measured by the cosine similarity.

We see that W2VPred correctly captured the domain-specifc meaning of "power": In *Natural Sciences* and *Engineering* & *Technology* the word is used in a physical context, e.g., in combination with generators which is the closest word in both categories; In *Social Sciences* and *Humanities* on the other hand, the nearest words are "powerful" and "control", which, in combination, indicates that it refers to "the ability to control something or someone".[6] The embedding trained by GloVe shows a

[6] `https://www.oxfordlearnersdictionaries.com/definition/english/power_1`

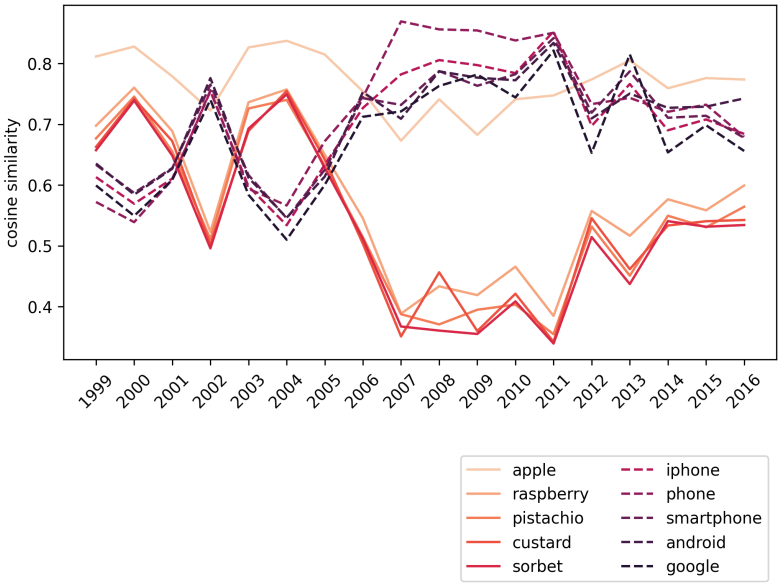

Figure 5.2: Evolution of the word *blackberry* in NYT. Nearest neighbors of the word *blackberry* have been selected in 2000 (blueish) and 2011 (reddish), and the embeddings have been computed with W2VPred. Cosine similarity between each neighboring word and *blackberry* is plotted over time, showing the shift in dominance between fruit and smartphone brand. The word *apple* also relates to both fruit and company, and therefore stays close during the entire time period.

very general meaning of power with no clear tendency towards a physical or political context, whereas Skip-Gram shows a tendency towards the physical meaning. We observed many similar examples, e.g., charge:electrical-legal, performance:quality-acting, resistance:physical-social, race:championship-ethnicity.

As another example in the NYT corpus, Figure 5.2 shows the evolution of the word *blackberry* which can either mean the fruit or the tech company. We selected two slices (2000 & 2012) with the largest pairwise distance for the *blackberry*, and chose the top-5 neighboring words from each year. The figure plots the cosine similarities between *blackberry* and the neighboring words. The time series shows how the word *blackberry* evolved from being mostly associated with the fruit towards associated with the company, and back to the fruit. This can be connected to the release of their smartphone in 2002 and the decrease in sales number after 2011.[7][8] The representation of the word *apple*, however, does change so much over the course of the years as it reflects both meanings, a fruit and a tech company.

5.3.5 Structure Prediction Performance

In this subsection, predicted structure of W2VPred is evaluated against the a priori given affinity matrix $D \in \mathbb{R}^{T \times T}$ (shown in the upper row of Figure 5.1) as the ground-truth. We report on recall@k averaged over all domains.

[7]See `https://www.businessinsider.com/blackberry-smartphone-rise-fall-mobile-failure-innovate-2019-11`.

[8]See `businessinsider.com/blackberry-phone-sales-decline-chart-2016-9`.

Dataset **Method**	NYT	WikiFos	WikiPhil
GloVe	67.22	51.66	36.67
Skip-Gram	71.11	54.59	26.67
CBOW	65.28	45.00	23.33
W2VPred (our)	81.67	62.50	23.33
Burrows'	55.56	22.92	6.67

Table 5.6: Recall@k for structure prediction performance evaluation with the prior structure (Figure 5.1 left) used as the ground-truth.

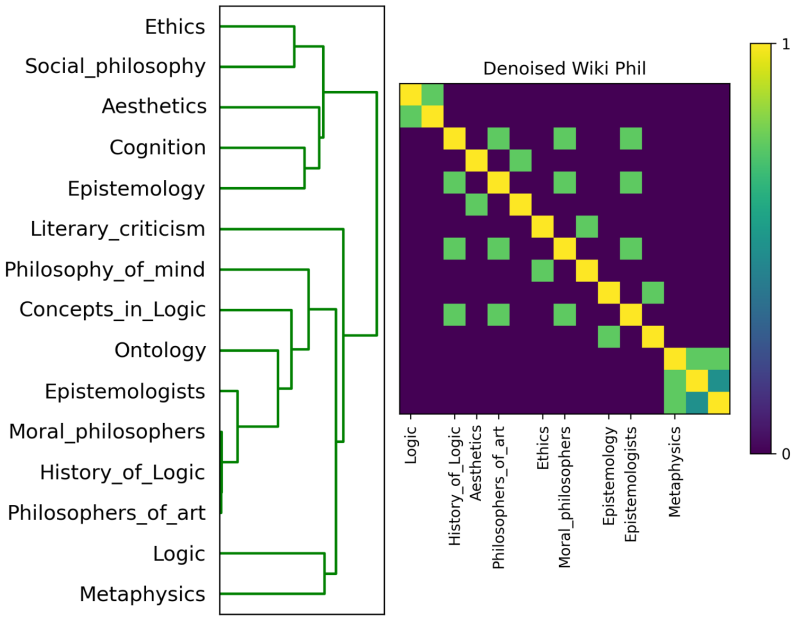

Figure 5.3: Left: Dendrogram for categories in WikiPhil learned by W2VPred based on the affinity matrix W. Right:Denoised Affinity matrix built from the learned structure by W2VPred. Newly formed Cluster includes *History of Logic*, *Moral Philosophers*, *Epistemologists*, and *Philosophers of Art*.

We compare the W2VPred method with Burrows' Delta J. Burrows 2002 and other baseline methods based on the GloVe, Skip-Gram, and CBOW embeddings. Burrows' Delta is a commonly used method in stylometrics to analyze the similarity between corpora, e.g., for identifying the authors of anonymously published documents. The baseline methods based on GloVe, Skip-Gram, and CBOW simply learn the domain-specific embeddings separately, and the distances between the slices are evaluated by Equation 5.4.

Table 5.6 shows recall@k (averaged over ten trials). As in the analogy tests, the best methods are in gray cell according to the Wilcoxon test. We see that W2VPred significantly outperforms the baseline methods for NYT and WikiFoS.

Structure Discovery by W2VDen The good performance of W2VPred on WikiPhil in Sub-Section 5.3.3 suggests that W2VPred has captured a suitable structure of WikiPhil. Here, we analyze the learned structure, and polish it with additional side information. Figure 5.3 (left) shows the dendrogram of categories in

WikiPhil obtained from the affinity matrix W learned by W2VPred. We see that the two pairs *Ethics-Social Philosophy* and *Cognition-Epistemology* are grouped together, and both pairs also belong to the same cluster in the original structure. We also see the grouping of *Epistemologists*, *Moral Philosophers*, *History of Logic* and *Philosophers of Art*. This was at first glance surprising because they belong to four different clusters in the prior structure. However, investigating articles revealed that this is a natural consequence from the fact that the articles in those categories are almost exclusively about biographies of philosophers, and are therefore written in a distinctive style compared to all other slices.
To confirm that the discovered structure captures the semantic subcorpora structure, we defined a new structure for WikiPhil, which is shown in Figure 5.3 (right), based on the findings above and also define a new structure for WikiFoS: A minor characteristic that we found in the structure of the prediction of W2VPred in comparison with the assumed structure is that the two subcorpora Humanities and Social Sciences and the two subcorpora Natural Sciences and Engineering are a bit closer than other combinations of subcorpora, which also intuitively makes sense. We connected the two sub-corpora by connecting their root node respectively and then apply W2VDen. The general analogy tests performance by W2VDen is given in Table 5.3. In WikiFoS, the improvement is only slightly significant for $n = 5$ and $n = 10$ and not significant for $n = 1$. This implies that the structure that we previously assumed for WikiFoS already works well. This shows that applying W2VDen is in fact a general purpose method that can be applied on any of the data sets but it is especially useful when there is a mismatch between the assumed structure and the structure predicted by W2VPred. In WikiPhil, we see that W2VDen further improves the performance by W2VPred, which already outperforms all other methods with a large margin. In the Appendix C.3, the change of performance due to the newly discovered structure has been additionally supported by investigating the correlation between the different evaluation metrics. There, it turned out that only for W2VConstr on WikiPhil, the Pearson correlation coefficient is even negative with -0.19.

5.3.6 Summary of the Evaluation on Benchmark Data

In this sub-section, we have shown a good performance of W2VConstr and W2VPred in terms of global and domain-specific embedding quality on news articles (NYT) and articles from Wikipedia (WikiFoS, WikiPhil).
We have also shown that W2VPred is able to extract the underlying subcorpora structure from NYT and WikiFoS. On the WikiPhil dataset, the following observations implied that the prior subcorpora structure, based on the Wikipedia's definition, was not suitable for analyzing semantic relations:

- Poor general analogy test performance by W2VConstr (Table 5.3),
- Low structure prediction performance by all methods (Table 5.6)
- Negative correlation between embedding accuracy and structure score.

		RW-STAN	MTurk-771	RG-65	WS-353-ALL	MTR-3k	WS-353-REL	MC-30	YP-13	WS-353-SIM	MTurk-287	SimVerb-350	SIMLEX-999	∑
NYT	GloVe	.36	.41	.46	.44	.50	.34	.50	.38	.53	.55	.15	.26	-
	Skip-Gr.	.50	.54	.51	.56	.63	.48	.56	.37	.64	.64	.22	.32	2
	CBOW	.55	.52	.40	.56	.60	.47	.40	.31	.62	.64	.25	.33	3
	DW2V	.51	.58	.55	.57	.68	.50	.58	.39	.64	.66	.23	.30	7
	W2VC	.52	.59	.53	.56	.68	.49	.54	.40	.64	.67	.23	.30	5
	W2VP	.53	.59	.54	.57	.68	.51	.54	.40	.64	.66	.23	.30	7
WikiFos	GloVe	.38	.58	.60	.57	.66	.52	.71	.33	.63	.44	.17	.28	1
	Skip-Gr.	.43	.60	.67	.65	.60	.59	.68	.41	.67	.62	.17	.28	3
	CBOW	.42	.59	.60	.63	.59	.57	.59	.43	.66	.64	.20	.29	-
	W2VC	.42	.62	.74	.62	.65	.55	.77	.44	.70	.68	.21	.29	8
	W2VP	.42	.62	.72	.62	.65	.55	.75	.43	.70	.69	.21	.29	7
WikiPhil	GloVe	.34	.49	.42	.53	.59	.51	.47	.31	.59	.44	.11	.23	-
	Skip-Gr.	.43	.55	.50	.59	.56	.54	.62	.40	.64	.63	.19	.27	4
	CBOW	.38	.49	.38	.57	.48	.51	.54	.39	.61	.61	.17	.25	-
	W2VD	.38	.59	.63	.59	.62	.55	.66	.43	.67	.66	.19	.27	9
	W2VP	.34	.58	.55	.58	.60	.54	.58	.40	.64	.65	.18	.26	1

Table 5.7: Correlation values from word similarity tests on different datasets (one per row). The best method and the methods that are not significantly outperformed by the best is marked with gray background, according to the Wilcoxon signed rank test for $\alpha = 0.05$. In this table, we use a shorter version of the method names (W2VC for W2VConstr, etc.).

	NYT	**WFos**	**WPhil**
GloVe	0.26	0.29	0.27
Skip-Gram	0.28	0.30	0.29
CBOW	0.29	0.31	0.29
DW2V	0.28	—	—
W2VConstr (our)	0.29	0.32	—
W2VDen (our)	—	—	0.30
W2VPred (our)	0.28	0.32	0.29

Table 5.8: QVEC results: correlation values of the aligned dimension between word embeddings and linguistic word vectors.

Accordingly, we analyzed the learned structure by W2VPred, and further refined it by *denoising* with human intervention. Specifically, we analyzed the dendrogram from Figure 5.3, and found that 4 categories are grouped together that we originally assumed to belong to 4 different clusters. We further validated our reasoning by applying W2VDen with the structure shown in Figure 5.3 resulting in the best embedding performance (see Table 5.3).

This procedure poses an opportunity to obtain good global and domain-specific embeddings and extract, or validate if given a priori, the underlying subcorpora structure by using W2VConstr and W2VPred. Namely, we first train W2VPred, and also W2VConstr if prior structure information is available. If both methods similarly improve the embeddings in comparison with the methods without using any structure information, we acknowledge that the prior structure is at least useful for word embedding performance. If W2VPred performs well, while W2VConstr performs poorly, we doubt that the given prior structure would be suitable, and update the learned structure by W2VPred. When no prior strucuture is given, we simply apply W2VPred to learn the structure.

We can furthermore refine the learned structure with side information, which results in a clean and human interpretable structure. Here W2VDen is used to validate the new structure, and to provide enhanced word embeddings. In the experiment on the WikiPhil dataset, the embeddings obtained this way significantly outperformed all other methods. The improved performance from W2VPred is probably due to the fewer degrees of freedom of W2VConstr, i.e., once we know a reasonable structure, the embeddings can be more accurately trained with the fixed affinity matrix.

5.4 Structures of German Authors' Literary Works

Now that the novel methods have been established and demonstrated to perform well in terms of embedding quality and predicted structure on the benchmark data sets, in this section it will be shown how the method can be applied on a digital humanities data set.

As shown, this method can be used to create embeddings that are subject to a certain structure but it can also be used to assess or question structure that is given a priori. As discussed before, the predicted outcome of the method can be compared against arbitrary affinity structures, typically given by the existing meta data of the subject corpus. The datasets chosen in the previous section were very distinct about which type of corpus structure is likely to be found, and still in one case, the WikiPhil dataset, the assessment after using the proposed methods suggested a different more fitting structure. In the case of digital humanities corpora it has been discussed at length in this thesis that there is typically very complex and diverse meta data available. This raises the question of what happens when multiple meta data dimensions are actually accounting for the found structure by W2VPred.

To this end, we applied W2VPred to high literature texts (*Belletristik*) from the lemmatized versions of DTA (German Text Archive, See *Deutsches Textarchiv. Grund-*

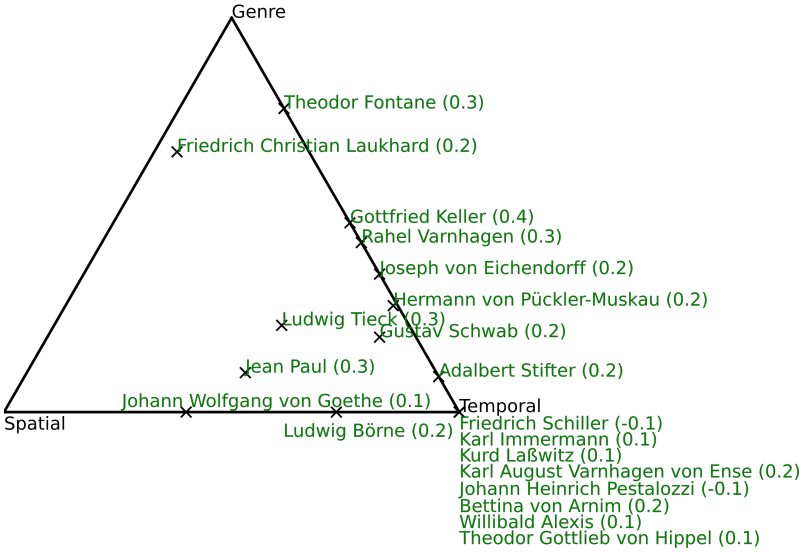

Figure 5.4: Author's points in a barycentric coordinates triangle denote the mixture of the prior knowledge that has the highest correlation (in parentheses) with the predicted structure of W2VPred. The correlation excludes the diagonal, meaning the correlation between the author itself.

lage für ein Referenzkorpus der neuhochdeutschen Sprache. 2022), a corpus selection that contains the 20 most represented authors in the DTA text collection for the period 1770-1900. For each author one sub-corpus was constituted and also we compiled three different candidates for structure from the meta data.
As a first measure of comparison, we extracted the year of publication as established by DTA, and secondly identified the place of work for each author[9] and finally categorized each publication into one of three genre categories (ego document, verse and fiction). Ego documents are texts written in the first person that document personal experience in their historical context. They include letters, diaries, memoirs and have gained momentum as a primary source in historical research and literary studies over the past decades.
In this experiment we want to compare the pairwise distance matrix that the proposed method predicted with the distance matrices that can be obtained by meta data available in the DTA corpus - the reference dimensions:

1. Temporal difference between authors. We collect the publication year for each title in the corpus and compute the average publication year for each author. The temporal distance between one author A_{t1} and another author A_{t2} is computed by $|A_{t1} - A_{t2}|$, the absolute difference of the average publication year.

2. Spatial difference between authors. We query the German Integrated Authority File for the authors' different work places and extract them as longitude and latitude coordinates on the earths surface. We compute the average coordinates for each author by converting the coordinates into cartesian system

[9]via the German Integrated Authority Files Service (GND) where available, adding missing data points manually.

and take the average on each dimension. Then, we convert the averages back into the latitude, longitude system. The spatial distance between two authors is computed by the geodesic distance as implemented in geopy.[10]

3. Genre difference between authors. We manually categorized each title in the corpus into one of the three categories ego document, verse and fiction. A genre representation for an author $A_g = (A_{g_{\text{ego}}}, A_{g_{\text{verse}}}, A_{g_{\text{fiction}}})$ is the relative frequency of the respective genre for that author. The distance between one author A_{g1} and another author A_{g2} is computed by $1 - \frac{A_{g1} \cdot A_{g2}}{||A_{g1}|| \cdot ||A_{g2}||}$, the cosine distance.

This resulted in three different affinity matrices between the authors that could all potentially be relevant and viable structures and could also be the ones W2VPred predicts.
After applying W2VPred to predict the connections between authors[11], it was to investigate which of the meta data structures can also be identified in the structures predicted by W2VPred, or in other words: Does it show in the textual representation that two authors published at similar times or in similar places, for example?

To find the connection between the representations and the structure label candidates, for each author, we correlated linear combinations of this (normalized) spatio-temporal-genre prior knowledge with the structure found by the proposed method which we show in Figure 5.4. For each author t, we denote the predicted distance to all other authors as $X_t \in \mathbb{R}^{T-1}$ where T is the number of authors. $Y_t \in \mathbb{R}^{(T-1)\times 3}$ denotes the distances from the author t to all other authors in the three meta data dimensions: space, time and genre. For the visualization we seek for the coefficients of the linear combination of Y that has the highest correlation with X. For this, Non-Negative Canonical Correlation Analysis with one component is applied. The MIFSR algorithm is used as described by Sigg et al. 2007.[12] The coefficients are normalized to comply with the sum-to-one constraint for projection on the 2d simplex.
For many authors, the strongest correlation occurs with a mostly temporal structure and fewer correlate strongest with the spatial or the genre model. Börne and Laukhard who have a similar spatial weight and thereby forming a spatial cluster, both resided in France at that time. The impact of French literature and culture on Laukhard and Börne's writing deserves attention, as suggested by this findings.
For Fontane, we do not observe a notable spatial proportion which is surprising because his sub-corpus mostly consists of ego documents describing the history and geography of the area surrounding Berlin, his workplace. However, in contrast to the other authors residing in Berlin, the style is a lot more similar to a travel story. In W2VPred's predicted structure, the closest neighbor of Fontane is, in fact, Pückler (with a distance of .052), who also wrote travel stories. Also, for Fontane, the

[10] `https://geopy.readthedocs.io/en/stable/`.
[11] With $\lambda = 512, \tau = 1024$ (same as WikiFoS).
[12] We use $\epsilon = .00001$.

unique mixture of poems, ego documents and fiction might play a central role.
In the case of Goethe, we also expected an overlap with the genre model as his genre distribution is as diverse as Fontane's. However, the maximum correlation at the (solely spatio-temporal) resulting point is relatively low and interestingly, the highest disagreement between W2VPred and the prior knowledge is between Schiller and Goethe. The spatio-temporal model represents a close proximity; however, in W2VPred's found structure, the two authors are much more distant. In this case, the spatio-temporal properties are not sufficient to fully characterize an author's writing and the genre distribution may be skewed due to the incomplete selection of works in the DTA and due to the limitations of the labeling scheme, as in the context of the 19th century, it is often difficult to distinguish between ego documents and fiction.
Nonetheless we want to stress the importance of the analysis where linguistic representation and structure, captured in W2VPred, is in line with these properties and also, where they disagree. Both agreement and disagreement between the prior knowledge and the linguistic representation found by W2VPred can help identifying the appropriate approach for a literary analysis of an author.

5.5 Summary & Conclusion

We proposed novel methods to capture domain-specific semantics, which is essential in many NLP and DH tasks: Word2Vec with Structure Constraint (W2VConstr) trains domain-specific word embeddings based on prior information on the affinity structure between subcorpora; Word2Vec with Structure Prediction (W2VPred) goes one step further and predicts the structure while learning domain-specific embeddings simultaneously. Both methods outperform baseline methods in benchmark experiments with respect to embedding quality and the structure prediction performance. Specifically, we showed that embeddings provided by the proposed methods are superior in terms of global and domain-specific analogy tests, word similarity tasks, and the QVEC evaluation, which is known to highly correlate with downstream performance. The predicted structure is more accurate than the baseline methods including Burrows' Delta. We also proposed and successfully demonstrated a procedure, Word2Vec with Denoised Structure Constraint (W2VDen), to cope with the case where the prior structure information is not suitable for enhancing embeddings, by using both W2VConstr and W2VPred. One of the main contributions of Word2Vec with Structure Prediction is that it finds word embeddings and structure between word embeddings at the same time. An underlying idea is that if the amount of data within the each sub-corpus varies Word2Vec with Structure Prediction makes the embedding matrices of the sub corpora with lesser data be supported by the data from other subcorpora with more data. The information for the word embeddings can flow between the different embeddings due to the neighborhood regularizer.
Sub corpora with lesser data will then lean on data from other sub corpora which

can have an influence on the predicted structure as their embeddings will become more similar. There are two forces competing in the optimization, the one that works toward building out a structure and the one that tries to equalize the embeddings. When laying more emphasis on improving embedding quality for sub corpora with lesser data it can happen that the predicted structure collapses. In general, the method is relatively robust toward changes of the hyper parameters but as one can see in Figure 5.1 (a), the 2006 sub-corpus that consists of fewer data is "falsely" predicted to be structurally closest to 2016.

Overall, we showed the benefits of the proposed methods, regardless of whether (reliable) structure information is given or not. Finally, we were able to demonstrate how to use W2VPred to gain insight into the relation between 19th century authors from the German Text Archive and also how to raise further research questions for high literature. This is especially due to the fact that the method is able to create individual textual representations for sub-corpora while retaining connections to others. This aims to be a middle ground between focussing on a specific subcorpus (again, methodologically speaking) without losing to much supplementary training data from other corpora. This is also due to the fact that it can be chosen between complementary methods W2VPred and W2VConstr whenever one wants to employ either confirmatory or explorative methods.

Part III

Modeling Phenomena of Textual Variants

Chapter 6

Alteration Types in Historical Manuscripts

In the two preceding parts the prerequisites were fulfilled to establish a level playing field of research between the two disciplines with respect to the data sets as well as with respect to the methodology. It has also been stressed already that the analysis of textual variants with the help of machine learning has potential use cases in different areas of inquiry within the humanities. In this part two studies from different areas are presented, text genetics (this chapter) and translation stylometry (Chapter 7). From the machine learning perspective, the main goal is to either adapt existing methods or develop novel methods that can take into account textual variants. From the humanities point of view, the goal is to test existing hypothesis about and explore so far unnoticed aspects of the corpora.

6.1 Introduction

For the analysis of text genetics, we chose a corpus that stands out by the level of detail that it has been prepared with. Even smallest alterations within the manuscripts have been annotated as such. The *Letters and Texts. Intellectual Berlin around 1800* edition is a digital scholarly edition of manuscripts by men and women writers of the late 18th and early 19th century (see Baillot 2022). The connection these writers have to the intellectual networks in the Prussian capital city are either direct (authors living and writing in Berlin) or indirect (editorial or epistolar relationship with Berlin-based intellectuals, see Baillot 2016). The originality of this digital scholarly edition is that it is neither author-centered nor genre-based, but presents different types of selected manuscripts that shed light on the intellectual activity of Berlin at the turn of the 18th to the 19th century. This editorial choice is presented at length in Baillot and Busch 2014, where light is also shed on the uniqueness of the Prussian capital city in the context of the period. While correspondences play a key role in the edition, they are considered as a part of the circulation of ideas that is at the core of the project, so that letters, and more generally ego documents, are complemented by drafts of either literary works (among which two major romantic texts), scholarly writings (one dissertation) or administrative documents (related to the de-

velopment of the Berlin University). A first editorial phase (2010-2016) allowed to publish manuscripts that cover different thematic areas and historical phases of the development of intellectual Berlin. They were selected based mainly on their scholarly relevance and on their accessibility (the publication policy of archives holding manuscripts having a major impact on their integration to a digital edition that displays a facsimile like this one does). Four main topics have emerged in this first phase: French, e.g. Huguenot Culture, Berlin University, Literary Romanticism and Women Writers. Depending on the topics, the letters published were complemented by other types of texts that document the circulation of ideas and of literary and scholarly works in the late 18th and early 19th century.
The edition can be browsed by theme, by author, by period, by holding institution, or by date. The single document can be displayed on one or two columns presenting at the user's choice a facsimile of the current manuscript page, a diplomatic transcription, a reading version, the metadata, the entities occurring on the page and the XML file corresponding to the document. In this first development phase, 248 documents and 17 authors were encoded and presented in BI. In Figure 6.1, a quantitative overview of the BI corpus is given, which consists of introductory figures for the whole corpus in terms of size, temporal span of documents and detailed information about individual authors.

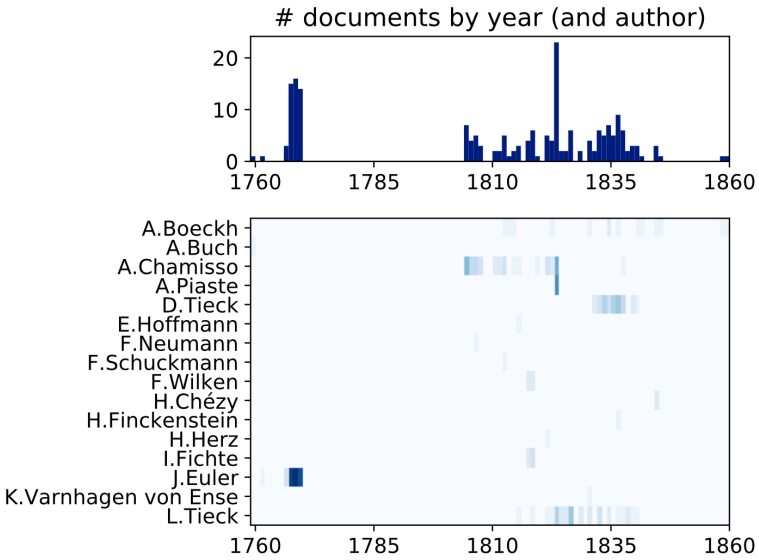

Figure 6.1: Temporal distribution of the creation of the documents is shown for the whole corpus (top) and for each individual author (bottom). In this subfigure (bottom), the number of documents created in each year is encoded in the intensity of the color.

A major novelty about the BI edition is that it combines genetic edition and entity annotation in order to gain insight in intellectual networks, on the actual editing process of manuscripts (of literary and scholarly works) and on the discourse about this editing process (letters – most letters are interestingly also partly transformed into literary works in their own right and subject to editing). The genetic encoding gives

precise information regarding deletions and additions in the manuscript text. The BI encoding guidelines make extensive use of the following specific sections of the TEI (P5)[1] guidelines additionally to the standard structure (Chapter 1–4): Manuscript Description (Chapter 10, `https://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html`), Representation of Primary Sources (Chapter 11, `https://www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html`) and Names, Dates, People, and Places (Chapter 13, `https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ND.html`), which offers the possibility to markup text alterations with tags such as `<add>` and `<del>`.

As already mentioned, the BI edition features annotations on the genesis of the documents (genetic edition), which, being a digital edition, are machine-readable. The core question we will address in the following is therefore whether machine learning models (Wainwright and M. I. Jordan 2008; Rasmussen and Williams 2006; Nakajima, Watanabe, and Sugiyama 2019) that analyze the alterations within the documents can be used to gain new insights into author, editor, and archivist practices, as well as practices of the intellectual societies in the document's creation time. The investigation of this question is only made possible by the meticulous (digital) annotation of the historical documents that provides previously unavailable enrichments and perspectives on the sources.

From the perspective of edition theory Ehrmann stresses that the importance of analyzing the alterations in manuscripts for literary studies and scholarly editing lies not only in the fact that they allow an insight into the author's writing process in the case of author-made changes, but also in the fact that they help identify the respective contribution in the case of co-authorships (Ehrmann 2016). The first question that arises when examining every alteration is the question of the underlying reason, be it for a minor correction of mistakes or a wide-ranging content-related alteration. This leads to the question of the originator of the alteration and, as Ehrmann stresses, whether the alteration is wanted by the author (Ehrmann 2016). In the specific case of an edition of correspondence, the intended readership is bound to change dramatically in the aftermath of publication. A letter that was originally written to a friend is made public to a large readership and in the process of preparation, the editor applies alterations to the original letter, most of the time with a correction phase on the original manuscript itself. This is the case for many manuscripts in the BI edition, commented on as follows by the editors:

> One characteristic of letters is that you generally are not the first one to read them when you discover them in an archive. Not only have they been addressed to a person or a group of persons in the first place [..], many of the letters we at least are working on have already been edited in the last centuries. But not in extenso, no: they have been abridged, overwritten, corrected according to the expectation of the audience in the time that they were edited. (Baillot and Busch 2015)

[1] The encoding guidelines can be found at `berliner-intellektuelle.eu/encoding-guidelines.pdf`.

Moreover, the novel machine learning method Alteration Latent Dirichlet Allocation (AlterLDA) presented here also offers new opportunities for many other areas of automated analysis of variants of sources, especially within the digital humanities. AlterLDA is based on the topic model latent Dirichlet allocation (LDA, see D. Blei, Ng, and M. Jordan 2003). The choice of using a topic model as the foundation for our method was also driven by the assertion that "Topic Modeling has proven immensely popular in Digital Humanities" (Schöch 2017) and it is therefore widely known and accepted in the field. LDA is particularly popular in the DH because it is suitable for explorative text analysis. With the automated compilation of word lists by LDA, new topics can be identified in large text corpora whose existence was previously unknown. In this context, it almost always forms the first analysis step on text data, but it can in fact also be used for non-textual data (Jelodar et al. 2019; Liu et al. 2016). In addition to LDA, which provides the identification of the overall relevant topics of the corpus to be examined and the specific topics of the individual documents of the corpus, AlterLDA is particularly concerned with the variants of the documents. The starting conditions for this work are as follows: from the point of view of edition theory, the question of document variants is of major importance, and this has not yet been sufficiently investigated with Distant Reading methods. From a methodological point of view, there is a very widespread topic model (LDA), which is already recognized and accepted practice in the digital humanities. In this section therefore the gap is closed by adapting LDA in order to model document variants.

The processual aspects of text genesis in the sources underlying the edition are thus highlighted and supported by the processual aspects of the edition itself. If a document in its past has already been prepared for publication by an editor, then his or her notes in the TEI-XML are annotated in the same way as when the editors of the BI edition leave notes: with the `<note>`-tag.

Parts may be deemed inappropriate for publishing to a broader readership at a certain place and time due to their political or religious context, or for revealing private information about a person or a group.

The application on the BI corpus is particularly interesting because the latter consists mainly of letters, which, especially around 1800 in Germany, exhibit a strong tension between public and private sphere. The framework presented here includes four methods that range from basic, well established, rule-based methods to a specialized, novel machine learning method (AlterLDA) that was developed for exactly this purpose. From a methodological point of view, this is a challenge for all disciplines involved, conceived as a scenario optimized so that all sides benefit from each other. Finally, the newly introduced method is also applied to discover alteration candidates in the documents that are not yet altered. These findings led to, and hopefully will continue to fuel, interesting discussions on parts of the edition that were unnoticed thus far. This approach is, again, methodologically motivated by "putting forth a reading" of a text in the sense of a radical transformation.[2]

[2] As it was discussed at length in Section 4.2 based on the theoretical concepts by Ramsay 2011.

For many applications, like the one presented in this thesis, we therefore rely on the evaluation by machine learning and humanities scholars, who can employ methods for interpreting and explaining machine learning models (Montavon, Samek, and K. Müller 2018; Samek, Montavon, Vedaldi, et al. 2019; Samek, Montavon, Lapuschkin, et al. 2021).

6.2 Methods

In this section, the machine learning methods for identifying the reason for a given alteration are presented, by first introducing the general data analysis pipeline. Then, we specify precise definitions of the most relevant concepts for alterations. After specifying the preprocessing steps, the novel AlterLDA model is introduced. It is designed to analyze the most interesting, yet most complex types of alterations. Before the methodologies of each step are explained in further detail, the definitions for the most important and most frequently used terms are given here.
Given an arbitrary version of a document, we define an alteration to be a local group of added and/or deleted symbols that is performed by the author of an alteration. Basically, any symbol appearing in the document could be regarded as a single addition, but the state of the manuscript at the time of the investigation often makes it impossible to identify beyond doubt which groups of symbols belong to a particular writing session. The same problem exists with deletions: Was the sentence completed first, or did the author pause in the middle and correct something before completing the sentence? In BI, additions and deletions are considered as such when they clearly stand out, for example when they are crossed out or written to the margin. Sometimes co-occurring additions and deletions are also referred to as replacements. The alteration may range from a single character to whole passages of the document and can even be a local group with non-altered symbols in between. An alteration author is a single person or institution that alters the document, possibly the primary author him or herself. The alteration author has an alteration reason for which he or she decides to alter the document. This is a very specific reason, for example “the alteration author thinks that a particular word is spelled differently” or “a real person which is referred to in the document may not want to be recognized by the readers, so this part is censored.”
Each alteration has a formal and content-related portion with varying emphasis. For example, if the author of an alteration changes the spelling of a single word this would not change the meaning of the document in most cases. On the contrary, adding multiple sentences to a document may change the content of the document significantly. Of course, whether an alteration is rather positioned on the form side or on the content side of the axis depends on the point of view of the recipient. Hence, the proposed method takes into account the formal changes of the document as well as the content-related changes. Smaller alterations tend to have a rather formal aspect, where longer alterations almost always are content-related.

The set of alterations can be broken down into different categories with respect to their alteration reason. One group of alterations is (1) the group of paratexts, for example archival notes, such as numberings or dates, or stamps and seals of the library or archival institution. Another group of alterations is (2) corrections of mistakes which consists in spelling alterations, grammatical changes and other corrections. (3) The third group contains stylistic alterations, for example replacing a token with its synonym or rearranging the word order. Of course, changing the word order is sometimes more than just a stylistic change, but one could e.g. begin a sentence with "Es bedarf daher [..]" as well as with "Daher bedarf es [..]" with very similar intentions. The last group of alterations which we call (4) content-related alterations incorporate alterations that either add new information to the document or suppress information that was present in the document before.
Figure 6.8 illustrates how the identification method works. All alterations are put into the analysis pipeline, and after the initial distinction between author alterations and non-author alterations, the four tests for different types are performed on each alteration. As an example, there are four alterations depicted in the illustration that are fed into the model. A detailed explanation for an identification of the three non-content-related types of modifications is given in the appendix. By elimination of all other possible categories, the remaining alterations are of the content-related category. There are still a variety of reasons in this category worthwhile to identify. Rather than the general category we aim for providing a distinct reason for each alteration. The fourth alteration which is marked in red is a longer deletion and a detailed facsimile is shown in Figure 6.8. It is performed with a pencil which is different from the primary ink of the letter. It deals with the author's sickness and with the sickness of the author's mother. The extent of the alteration already indicates that this is not a correction of a mistake and since the part that is deleted is not replaced by anything else, it can be assumed that this alteration changes the amount of information provided. It is thus to be classified as a content-related alteration. At this point it is still to be identified for which specific reason the document has been altered. With AlterLDA, the alteration is assigned to one of a set of candidate reasons as a final step, in this case Sickness-reason.

6.3 Related Work

We convey a generative topic model, that is based on Latent Dirichlet Allocation (D. Blei, Ng, and M. Jordan 2003) and that is able to take into account the structural information of alterations. LDA is a widely used topic model that extends Latent Semantic Indexing (Deerwester et al. 1990) which is capable of assigning a distribution of topics to a document instead of only a single topic. LDA takes advantage of the fact that a text is organized in documents. This structural information is the reason for LDA to function. Based on this structure of documents, LDA can learn which words tend to co-occur and thus have a relation. Words that often occur with

each other form a topic. In this context, a topic is merely a distribution of word frequencies.
There exists a wide range of topic models that customize LDA by taking into account additional structural information. To replace the bag-of-words approach by introducing structural information about the word order is a major field of LDA research (Rosen-Zvi et al. 2004; Wallach 2006; Gruber, Weiss, and Rosen-Zvi 2007). In addition, there is a broad research community that addresses the recognition and arrangement of hierarchies of topics (D. M. Blei, Griffiths, and M. I. Jordan 2010; Paisley et al. 2015). LDA has also been modified to work with graph-structured documents (Xuan et al. 2015). However, we are not aware of any literature that shows how to model alteration reasons in a corpus of natural language. Therefore, in this chapter an important contribution is made to close this gap, i.e. to provide the literary scholarly community with a novel method and to open up another field of application for the machine learning community. In Figure 6.2, the LDA

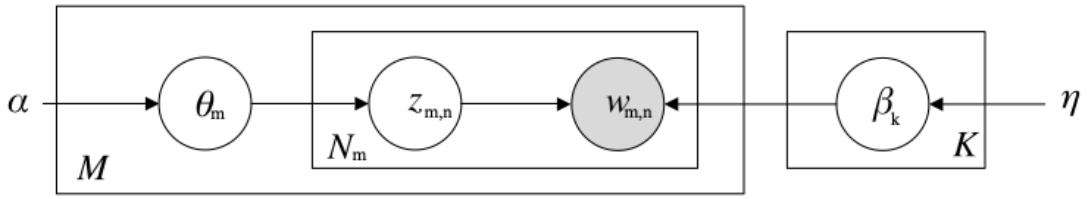

Figure 6.2: Graphical representation of the LDA model. The plate notation visualizes the generative process of a probabilistic model by following the directions of the arrows. Given α and β, one initially draws β, a distribution over words for each topic and θ, a topic distribution for each document. Then, for each token within a document, one draws a topic assignment and only then (because w has input arrows from both, β and z) one can draw w from the topic in β, that was assigned in z.

model is shown in plate notation. An overview over the used symbols is given in Appendix D.1. The plate notation shows the graphical representation of the LDA model. An open circle denotes a model variable and a shaded circle denotes an observed variable. Symbols without circle denote a hyper parameter. A rectangle indicates repetitions of the included variables. In this model, β represents the topic histograms, θ represents the topic mixture for each document, z represents the topic assignment for each token position and w denotes the token itself. LDA has no notion of the order of words within a document, which is referred to in the literature as a "bag-of-words" for each document.

6.4 Alteration Latent Dirichlet Allocation

In Figure 6.3, the AlterLDA model is described in plate notation. The upper part is standard LDA whereas the lower right part contains the newly introduced variables

to model alterations.
In standard LDA, the observed variable (the input) is just the words within each document. In the AlterLDA setting, the additional structural information about the alteration of each word is provided as input. With that, the AlterLDA model tries to infer the tendency for each topic to be an alteration topic.

The generative model detects reasons by taking into account all text, inside and

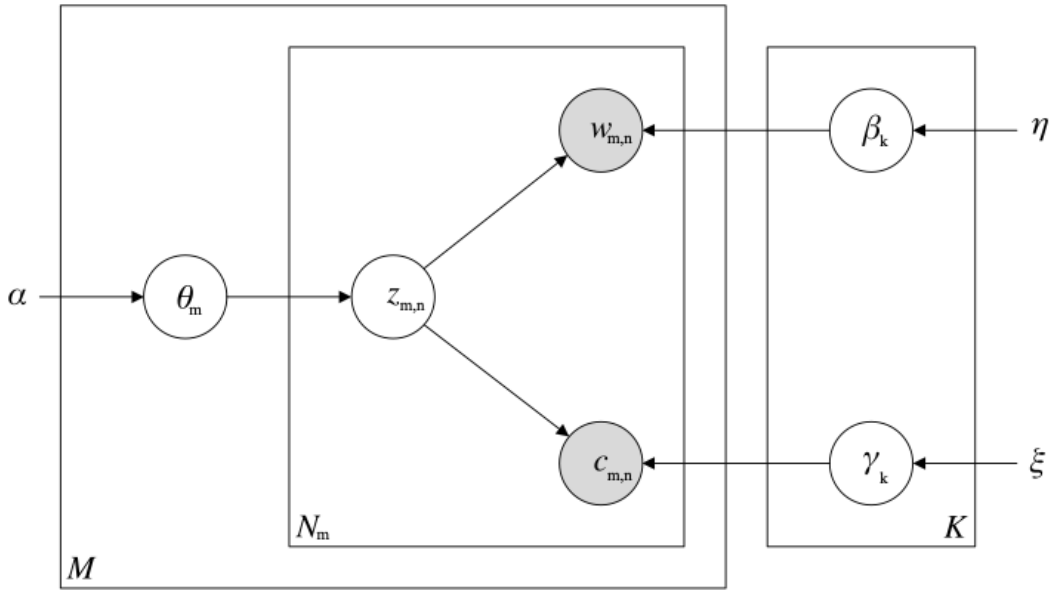

Figure 6.3: Plate notation of the new AlterLDA generative model. Newly introduced is the lower branch with variables c, γ, and ξ that deal with alterations. There exist M documents with N_m tokens each. Also, there exist K topics and for each topic, there exist a tendency for it to be a reason for alteration (γ).

outside the alterations. From alterations that were gone through manually, we expect to see alteration suggestions that mainly relate to the privacy of a person, political or religious topics may appear as well. In order to make the model description as clear as possible, we try to keep the mathematical formulations to a minimum. Therefore, we only include an explanation of the symbols used (see appendix), a graphical representation and the derivation of how the model can be algorithmically captured using a Collapsed Gibbs Sampler. Similar to LDA, in AlterLDA there exists no feasible algorithm to compute the posterior distribution of the latent variables. Instead, approximate methods need to be applied to find a solution in reasonable time.
A Collapsed Gibbs Sampler is one of the possible approaches to find an approximate solution to the objective. Generally, a Gibbs Sampler iteratively samples the configuration of a specific latent variable based on the current configuration of all other model variables. An introduction on Gibbs Sampling LDA is presented by Carpenter 2010. This algorithm can also be understood as an instance of a Markov Chain, a constrained iterative probabilistic model itself, where the current state only depends on the previous. From this perspective, the stationarity of the Markov Chain represents the solution of the Gibbs Sampler. The source code of our implementation

of the Collapsed Gibbs Sampler for the AlterLDA model is publicly available.[3] In the Appendix D.1, the derivation of the Collapsed Gibbs Sampler for the AlterLDA model is given.

6.5 Results

In this section, we present three experiment settings which mainly differ in the splitting between training and test data. As shown in Figure 6.4, three settings are chosen, S1 as a straightforward explorative demonstration, S3 to comply with the methodological standards of data splitting for the performance report, as well as S2 for offering additional explorative results specific for this data set. We will first present the evaluation results that investigate the performance of AlterLDA on the given data set and afterwards present explorative results that will be reconciled with expert knowledge. Apart from these experiments on the BI data set, the first experiments were performed on synthetic data, some results from these experiments are listed in the appendix.

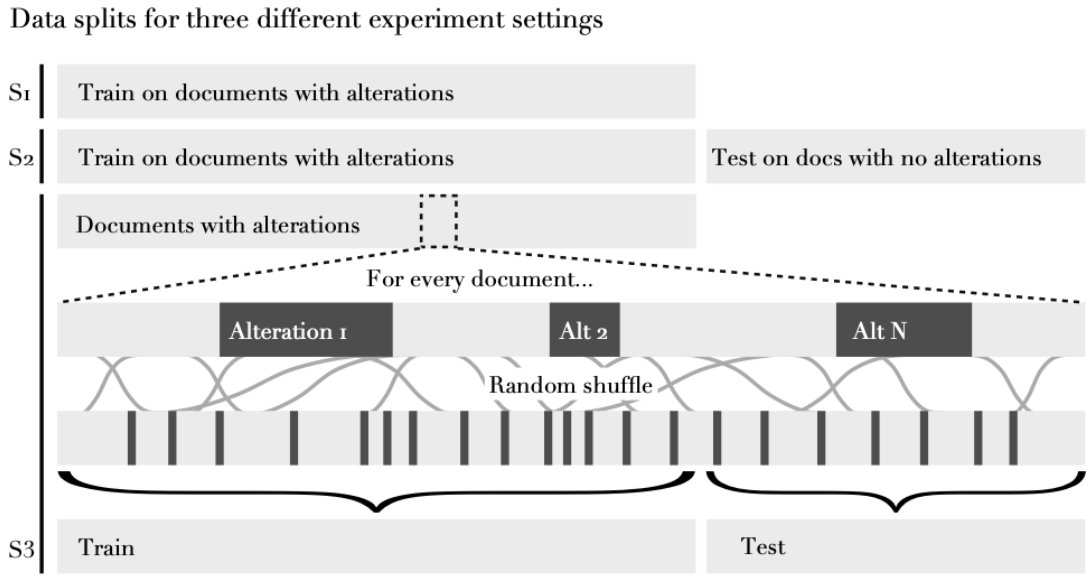

Figure 6.4: Visualization of the data splitting setup for all settings. For each experiment a different data setting is used. The different Settings are shown in the leftmost column (S1, S2, S3). Within each setting, the row at the bottom depicts the final setting of the data. Setting 1 (only one row) does not require test sets, Setting 2 (only one row) aims at finding alteration candidates in texts with no alterations. For Setting 3, the process of creating the setting is depicted in multiple rows. First, only documents that contain alterations are chosen. Then, each individual document is shuffled and split into a training and a test part.

[3] See `gitlab.tubit.tu-berlin.de/david.lassner/shipping_alterLDA`.

6.5.1 Performance Evaluation

In this experiment, in which AlterLDA is applied to the entire training data, it is to be determined whether the model in principle delivers plausible results. It will be verified whether γ finds a meaningful topic composition that represents sensitive topics. This means that alteration topics may be a convolution of private and maybe political and religious matters.
With AlterLDA, various parameters must be set which influence the outcome. These are the same parameters as for LDA: Number of topics and the Dirichlet prior for the topic distribution η and the topic mixture α. There is also another parameter, the Dirichlet prior for the alteration tendency ξ. The default value for a Dirichlet prior is 1, but it can take any value greater than 0. The smaller the value, the more the variable tends to be focussed on single values, the larger the value, the more different values are considered. Using the topic mixture as an example, a small α would mean that LDA is looking for a solution where each document consists of only a few topics, a large α finds a solution with mixtures of many topics. AlterLDA is initialized in this setting with $\alpha = (.1, .1, \ldots)$, $\eta = (.1, .1, \ldots)$ and $\xi = (.05, .05, \ldots)$ as well as K=10. We choose small in this setting to create a sparse so that AlterLDA only learns one alteration topic.
The resulting topic learned in this naive approach as alteration-sensitive is visualized as a word cloud in Figure 6.5. It is very difficult to put a single label on this "topic". The most probable words are strongly influenced by global word frequencies, the strongest four words describe it: "Sie Ich Brief schreiben". This does not come as a surprise since the corpus consists mainly of letters. However, it is also possible to find terms from any subject area that was suspected in advance of being altered: Sickness terms are for example "Operation", "Bett", "fürchten", Financial terms are e.g. "Geschäft", "Geld" and regarding Love Story there is for example "lieb" and "schön".
Beforehand, we assumed to also find political, religious topics but these do not appear in the naive setting whereas diverse private topics do occur.

As visualized in Figure 6.4, the BI corpus consists of documents with alterations and documents without alterations. If we want to measure the performance of AlterLDA in predicting the tendency for alteration, documents with alterations are much more helpful.
In Setting 3, we only use documents with alterations to produce the training and test set. We split every document individually into training and test set after shuffling to increase the chance that alterations are present in both sets (K.-R. Müller et al. 2001). After training, we use the topic mixture θ of the corresponding document.
In this setting, AlterLDA is initialized with $\alpha = (1, 1, \ldots)$, $\eta = (1, 1, \ldots)$ and $\xi = (1, 1, \ldots)$ as well as = 20. We explicitly chose the Dirichlet priors all equal to 1 as this can be considered the default. To allow for more topic diversity, we chose the number of topics a little bit higher than in the naive setting. This parameter combination will be used throughout the rest of the section.
The performances on the total test set as well as for each individual author are shown

Figure 6.5: The strongest words of the topic that has a high alteration tendency after training AlterLDA in the naive setting (Setting 1). The stronger the word, the larger the font. Strongest words are very general words on the topic of letters, words from the Financial, Sickness and Love Story are also weakly present.

in Table 6.1. The performance varies considerably across different authors where D. Tieck, L. Tieck and Ad. Chamisso work well above chance level, the performance for Hoffmann and especially H. Finckenstein is weaker. In case of H. Finckenstein, this likely due to the fact that in the corpus there is only a single letter and except for the salutation, the whole letter has been struck through. This also explains the discrepancy between AUC and balanced accuracy.
For E.T.A. Hoffmann, there is also only one document in the corpus, but it presents two specialties. It is considerably longer than most documents in the corpus: it is not a letter, but the novella Der Sandmann (the sandman). The larger size and the differing properties due to the genre seem to trade off to a slightly better performance than in the case of H. Finckenstein. We thus argue that the performance of AlterLDA depends on the size of the training set and on the homogeneity of the documents.
The results of this setting are not meant produce new domain insights as it only aims at reproducing the alteration tendencies of already altered documents. However, screening performance difference across viewpoints such as authors still reveals properties of the underlying data set.

6.5.2 Explorative Analysis

In the explorative experiment (S2), the corpus is divided into two parts: On the one hand, all documents that contain changes and, on the other hand, all documents

Grouping	Balanced Accuracy	Area under ROC
Adelbert von Chamisso	.60	.57
Henriette v. Finckenstein	.38	.07
Immanuel v. Fichte	.49	.64
E.T.A. Hoffmann	.5	.53
Dorothea Tieck	.61	.65
Ludwig Tieck	.69	.65
Total	.67	.66

Table 6.1: Test set performance for documents with alterations. The test set is grouped by author and two performance measures are given. Balanced Accuracy and Area under Receiver Operating Characteristic.

that do not contain any changes. The aim of the experiment is to train the model on the part of the corpus that contains changes and then let the model suggest which parts of the unchanged corpus may be changed in a similar way. There may be different reasons why some documents contain alterations and others do not. Assuming that all documents were reviewed by the same person and that person was also so diligent that he or she did not overlook a single passage, then AlterLDA should at best-case scenario not propose an additional passage to be altered. We assume in this experiment that either not all documents have been reviewed for the same criteria or that relevant positions have been overlooked.

In Table 6.2 the counts of the positions in documents with no alterations that have been suggested by the method are displayed for each author/editor. The rows in the table are sorted by the total amount of suggested alterations. Interestingly, the authors with many suggested alterations are not necessarily the ones that have a large share of total tokens of the corpus. In the case of Euler and von Buch, this is due to the fact that their documents are mostly in French, whereas the AlterLDA model in this case is primarily trained on German texts. For Boeckh, this is mainly due to the fact that the corpus encompasses only a few yet long documents and consequently there are not many documents present in the test set. Of course, there are other reasons for each author's ratio of corpus portion and number of suggested alterations. A person that altered all positions in the training set also diligently edited all documents in the test set and simply did not find any position that should be altered for the same reason: That the method did not find the respective amount in the test set can either mean that it was not able to find the right positions or that there were none.
For further analysis, we will ignore the texts by J. A. Euler and A. F. Buch and focus on the other four authors for which AlterLDA suggested most alterations. As said, the texts by J. A. Euler and A. F. Buch were mainly written in French which influenced the number of suggested alterations. In Table 6.3, the most common words that were suggested to be altered for individual authors are listed. For all authors except A. Boeckh, the majority of words seem to relate to the overall letter topic, however for D. Tieck the keyword "Krankheit" Sickness appears. For Ad.

Author	Suggested alterations
Immanuel Hermann von Fichte	3
Karl August Varnhagen von Ense	16
Friedrich Wilhelm Neumann	54
Helmina von Chézy	59
Adelheid Reinbold	73
Henriette Herz	76
Friedrich von Schuckmann	118
Antonie von Chamisso	258
Friedrich Wilken	340
Ludwig Tieck	389
August Boeckh	540
Adolf Friedrich von Buch	907
Dorothea Tieck	929
Adelbert von Chamisso	1075
Jean Albert Euler	2558

Table 6.2: The table shows the number of suggested alterations for different authors/editors. Only documents that were not truly altered are included. The number of suggested alterations represents the number of positions in documents that the method suggests to alter. Euler and Buch mainly wrote in French, which influenced the prediction a lot for these authors. Chamisso wrote in German although his mother tongue is French. Therefore there may be many minor mistakes that are suggested to be altered. Dorothea Tieck wrote about her mother's sickness which the method recognized as a sensible topic and therefore as a reason for alteration.

Chamisso, words like e.g. "schön" and "begehren" can be observed that may relate to the topic Love Story. In the case of L. Tieck, a distinct convoluted alteration topic is not immediately conceivable. In the case of A. Boeckh, the topic of the documents in the corpus are mostly academia-related.

For a better understanding of alteration suggestions, a closer look into the individual authors is provided. D. Tieck's documents reveal a sequence of letters that she wrote to F. Uechtritz in the years between 1831 and 1840. In the letters she repeatedly mentions her mother's sickness until her death in February 1837. Later, in March 1837, the father of F. Uechtritz passed away as well, D. Tieck writes about this in Letter 28.
By manually reviewing this series of letters, the editors of the BI edition agreed in many cases with the classification of the alterLDA model that the sickness of D. Tieck's mother plays a role for the alteration tendencies of the documents. In some cases, however, human experts and the model disagreed about the reason for alteration. The following excerpt from letter 12 is identified by our proposed method as a stylistic alteration.

> Meine arme Mutter ~~leidet schon seit längerer Zeit an Unter=~~
> ~~leibsbeschwerden, der Arzt sagt es seyen~~
> ~~Verhärtungen und Anschwellungen der Drü=~~

Author	**25 most common suggested alteration words** (descending order)
Dorothea Tieck	Sie, Brief, schreiben, Ich, schön, gewiß, denken, einig, lesen, all, gleichen, Düsseldorf, Agnes, Krankheit, Freund, kennen, erhalten, Arbeit, Dresden, halten, weiß, Die, Leben, Berlin, Lüttichau
Adelbert v. Chamisso	Brief, schreiben, Ich, de, Die, weiß, kennen, all, wissen, schön, 4, Freund, denken, Sie, gleichen, halten, 3, neu, ton, erhalten, begehren, bleiben, einig, lesen, Sache
Ludwig Tieck	Sie, Freund, Ich, Geist, Brief, Tieck, Dresden, Von, Ihr, umarmen, Juli, halten, sogleich, erleben, Die, schwach, schweigen, sprechen, Mich, Herrn, Vergnügen, fordern, Masse, gleichen, eintreten
August Boeck	Mitglied, Seminar, Sie, Prämie, Fichte, erhalten, Arbeit, 1813, 2, Verfasser, 1812, Übung, hiesig, 4, Fähigkeit, welch, außerordentlich, Nummer, zahlen, Prüfung, Wernike, Anstalt, anfangen, Gedicht, Studiosus

Table 6.3: Most common words that were suggested to be altered in texts from individual authors.

> sen, sie hat schon seit längerer Zeit viel zu leiden, ~~braucht schon seit 3 Monathen,~~ trinkt seit 4 Wochen hier
> Karlsbad, und ~~alles~~ bis jetzt ohne den
> mindesten Erfolg.
> (BI, Dorothea Tieck to v. Uechtritz, Letter 12, p. 2)

And one can argue that this is actually a stylistic alteration, because the information about the mother's sickness is preserved after the alteration. However, the detail that her mother has pelvic complaints is suppressed in the second version – this discrepancy in detail can be decisive for classification as a content-related alteration. In Figure 6.6, the number of suggested alterations (for the test set) and the number of actual alterations (for the training set) are displayed for each document by Ad. Chamisso. Most of the letters are addressed to L. de La Foye (Ad. Chamisso's best friend), some are addressed to Antonie von Chamisso (Ad. Chamisso's wife). For each of the addressees, the letters are ordered by date. There are two letters (letter 10 and 11) which stand out significantly with regard to the number of alterations, letter 10 actually encompasses a large number of alterations, whereas letter 11 is part of the test set and thus does not have any (content-related) alterations. The AlterLDA model suggests an almost equally high number of alterations for letter 11, presumably because it consists of topics that the AlterLDA model estimates to be altered accordingly - this shows that the AlterLDA model captures subtle changes of topics by the same author.

In Figure 6.7, the correspondence from L. Tieck to F. Raumer that is depicted ranges from years 1815 to 1840. The left panel showing the number of letters that were sent during that year grouped by whether they contain content-related alterations. The right panel shows the number of tokens that were altered (blue)

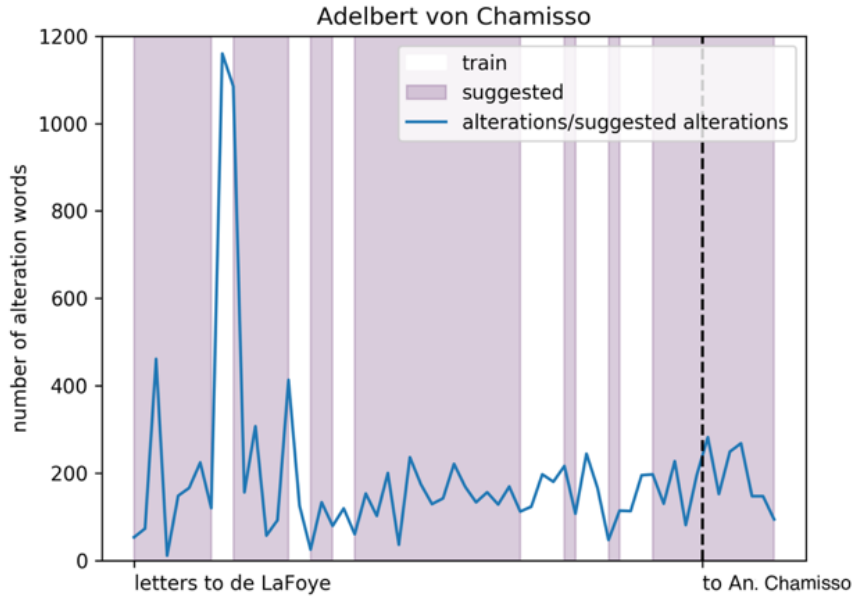

Figure 6.6: The letters are divided into documents containing content-related alterations (white background) and documents without content-related alterations (purple background). The AlterLDA model is trained on the white part and predicts possible alterations on the purple part, so the blue line shows the number of real alterations and suggested alterations, depending on the background. Left of the dotted separator, we find letters addressed to L. de La Foye, on the right side letters addressed to An. Chamisso. There are two consecutive outliers with significantly higher numbers of alteration words. One is part of the training set, one is part of the test set, the temporal proximity may indicate a content-related proximity that the model was able to capture.

and the number of tokens that AlterLDA suggests to be altered (orange).

Just comparing the blue bars of the two panels reveals that despite the fact that in 1836 there was only one letter written, there occurs the third-highest number of altered tokens. By examining the letter, it turns out that Tieck wrote about his financial problems and his plans to sell his book collection to the Count Yorck von Wartenburg.[4]

Referring back to Table 6.3, Financial terms are not present in the most common alteration suggestions for L. Tieck. This could indicate that the person editing L. Tieck's letters did not miss parts that refer to this financial struggle.

When also considering the suggestions by AlterLDA (the orange bars of the panel on the right), one letter from 1838 draws the most attention just by the sheer number of suggested alterations. In this letter, L. Tieck refers to disputes between the Catholic Church and the Prussian state at that time.[5] By arguing about this

[4] "[..] Tieck plante aus finanzieller Bedrängnis heraus den Verkauf seiner Bibliothek an den Grafen Yorck von Wartenburg[..]" BI, comment by Johanna Preusse in letter from Ludwig Tieck to Friedrich von Raumer (Dresden, 11. November 1836).

[5] "Kontext der von Tieck angedeuteten Vorgänge waren Machtstreitigkeiten zwischen der katholischen Kirche und dem preußischen Staat. [..]" BI, comment by Johanna Preusse in letter from Ludwig Tieck to Friedrich von Raumer (Dresden, 27. März 1838).

political controversy, L. Tieck chooses his wording in such a way that AlterLDA suggests alterations. This could indicate that across the training corpus there might be the same tendency to alter parts of the documents that deal with a political controversy. The fact that AlterLDA highlights a document containing a mixture of political and religious topics supports the hypothesis that the alterations in the BI corpus do not only consist of privacy matters, but also involve a wider political dimension. This result confirms and gives a novel dimension to the assertion that letters as a text genre evolve, especially in the German context of the 1800s, at the interface between private and public matters. In that sense, the role played by alterations aiming at balancing private and public dimensions is central and needs to be further delved into. AlterLDA provides a systematic approach to this major issue in literary studies.

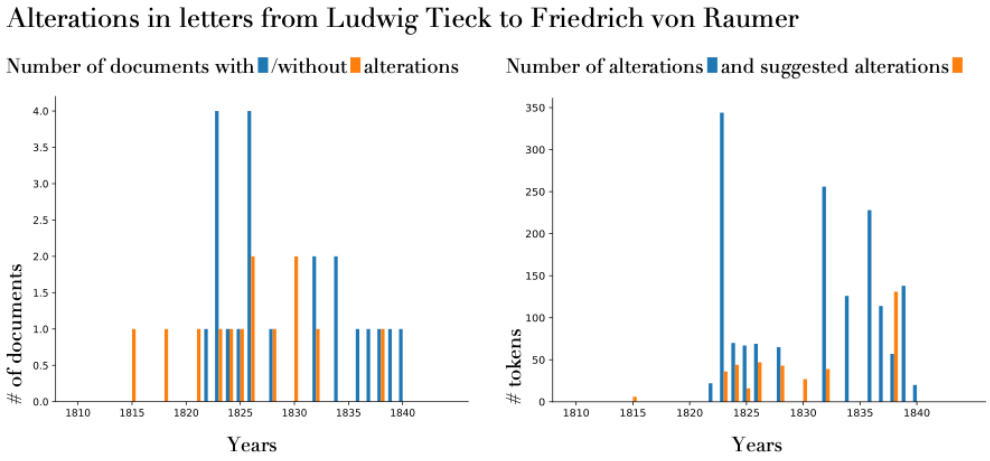

Figure 6.7: Left panel shows the timeline with counts of letters from Ludwig Tieck to Friedrich von Raumer. The right panel shows the number of tokens that were altered (blue) and the number of suggested tokens (orange). The comparison between the number of letters and the number of alterations for each year shows that there are times where the letters were altered more (e.g. 1836). The letters with suggested alterations deal with financial, political and religuous topics.

6.6 Summary & Conclusion

This Chapter has presented a general framework for analyzing alterations in historical documents, ranging from simple error corrections to stylistic changes and even to content-related alterations. In addition to established methods such as regular expressions, string distances and vector space comparison, a new probabilistic model for the modeling of reasons for alterations has been introduced (AlterLDA).
This work contributes to the understanding of text genesis, as it provides insight into the layers of changes in documents. It also offers a quantitative way of evaluating which topics are at what times prone to be altered and are therefore sensitive.
Besides the aforementioned complex annotation structure, from a machine learning

point of view, the BI data set posed special challenges because, on the one hand, the data set is very small and, on the other hand, it comprises several languages, which also differ greatly from the ones used today. Nonetheless, AlterLDA was able to confidently find alterations on unseen, labelled data. Exploratively, the method was able to find characteristics on unseen, unlabeled data that in many cases match the expert analysis. The method hence proves to be useful to draw the human reader's attention to specific parts of the large corpus that may otherwise be unnoticed, and by doing so serves as an example of how a machine learning method may assist a scholar as a collaborating reader and a potential collaborating editor.

With regard to the editorial issues first presented here, it is to be noted that the machine-readability of the BI edition makes it possible to serve problem-specific, individualised editions tailored to the research question of a reader/scholar. This project showcases how machine learning methods radically transform the way in which scholars engage historical documents, by taking advantage of the quality of deeply-annotated data: editorial and machine learning expertise can be brought together to explore in depth humanities research questions.

Figure 6.8: Flow chart of machine learning pipeline with four example alterations. The stream of documents is analyzed in four steps that identify different reasons of alterations as depicted in the panel at the top. In the panel at the bottom, the details of the individual alterations are presented. Each alteration has a unique appearance and unique characteristics, like the type of ink and the way in which it fits into the surrounding script. The presented preview of the facsimiles are shown in greater detail in the Appendix D.2.

Chapter 7

Translatorship Attribution and Translator Style

7.1 Introduction

In this Chapter, a famous German Shakespeare translation from the early nineteenth century will be analyzed. There exist numerous translations of Shakespeare's plays into German and each translation of a play can be seen as a variant of the original – in a different language. Since it was first printed, the translation of Shakespeare's plays edited by August Wilhelm Schlegel and Ludwig Tieck has been re-edited many times (Shakespeare 1833). A major reference in the first half of the 19th century, it is still regarded as a groundbreaking translation and referred to today. While there is little doubt that Schlegel translated the first edited plays, L. Tieck did not work out the edition of the final volumes by himself, but delegated the main translation work to his daughter Dorothea Tieck and Wolf Heinrich Graf von Baudissin (Paulin 1998). Although his daughter played a major role in the translation project, in the foreword of the first edition Ludwig Tieck only mentions that "a friend" helped him with the translation.[1] This Section investigates the contribution of the actors involved in this joint translation project. Machine Learning methods are used to analyse the English plays and their corresponding German translations in order to gain quantitative insights into what may seem a peculiar writing setting, but was quite usual in the context of the 19th century. The method proposed here is hence likely to improve our understanding of co-creation conditions in the 19th century at large.

In our setting, we first show which plays are translated by whom, based on D. Tieck's statement of the repartition of the plays (Uechtritz, Erinnerungen p. 173 and p. 177 - see Figure 1). Since the manuscript of the raw translation is now lost, the sole material this study can base its analysis on is the Shakespeare edition and the first German edition. We have no material traces allowing to easily discriminate between what D. Tieck translated, what W. Baudissin translated, and what L. Tieck corrected in the translations. We investigate two questions: firstly, it is still unclear

[1] The German "ein Freund" (a friend) specifies the male gender, for a female fried, one would write "eine Freundin".

which scenes of the plays in which D. Tieck and W. Baudissin collaborated have actually been translated by whom, so that a first goal consists in defining the roles and tasks of the three translation partners. The second point of interest is to shed light on the daughter-father relationship between D. Tieck and L. Tieck.
In general, it is not only very challenging to compare literary works across different languages, but also in the sense that, in contrast to authorship attribution, translators are aiming at preserving the style of the original text – the traces of the translators should therefore be even harder to identify. By offering methods to meet this challenge, this section presents a novel approach to use methods such as Burrows' delta in the multilingual context, to compare translation styles and attribute translators.

7.2 Related Work

Methods of stylometry are widely used to identify the author of a literary text (J. Burrows 2002; Argamon 2008). The idea is that an author expresses a unique style and that stylometry is able to capture the style and represent it as a kind of finger print of the author. Stylometry also been used to identify the translator of a literary text. – There is the work by Hoover 2019 that shows that identifying the author is even easier in the translations than in the original – identifying the translator, however, is only possible with hand-selected features (e.g. strong regularization) that may not generalize well. The fact that the author of the original text can be more precisely identified in the translation may be explained by the fact that a translator, especially when translating literature, is explicitly working with the author's style in order to mimic it. That way, the original author's style, that, in the original may sometimes be more implicit, is revealed more direct by the act of translation.
There is also the work by Burrows on poetry translation that shows that if a person translates and writes their own poems, it is often possible to identify the author style with a model that was only trained on the translations (J. F. Burrows 2002). – this is, of course only possible in the case where the translator allows – conscious or unconscious – their own style to be added to the poem that they translate.
The style of a translator as a literary style has been theorized by Bertin in his theory of translation: The act of translation is the transformation of a text from a source language into a target language. The more adopted the translation is to the target language, the more it is estranged with the original – the more characteristic is kept from the source language, the more it is estranged with the reader in the target language. Where the translator places themselves on this axis can be described as expression of translator style.
Even if this axis is a way for the translator to express their style, translated texts in general differ from texts that originate in the target language. This phenomenon is called translationese and it was already described by (Baker, Francis, and Tognini-Bonelli 1993; Gellerstam 1986). One widely known feature of translationese is that sentences in the process of translation, become shorter.

Stylometric investigations of collaborative translations to identify translators has already been analyzed by Rybicki and Heydel 2013, who could show that a Nearest Neighbors Classifier on Burrows delta features was able to distinguish between the different translators of novels by Virginia Woolf into Polish.
All the aforementioned works on translatorship attribution show that a major difficulty in translator identification and translator stylometry is that there are many confounding variables.
In the following, we want to give a set of examples which variables may confound the analysis of translator style:
First, there is the original author that always exist in a translation setting. Every translator has to position themselves toward the author. Without any knowledge about the original author, a setting where we only have samples of (Author A, Translator B) and (Author C, Translator D) it will be very hard to separate features of the author from features of the translator.
Second, there is the genre of the original text – for example, whether the original is a novel or a poem has a strong influence on how the translator translates. Again, settings of (Genre A, Translator B) and (Genre C, Translator D) will pose a difficult problem to isolate features of translators and features of genre.
These two were examples of confounding variables that are temporally prior to the act of translation. Additionally, one has also to deal with confounding factors that influence the data sources after the act of translation, such as a copy editor who prepares the translation for publication. This is of major importance in a collaborative translation setting as the one we are describing in this section.
Certain features that could have distinguished translators may have been normalized by an editor. Or, at the same time certain features may also be introduced by the editor and could therefore falsely be attributed to a translator.
Finally, another confounding variable can be alterations in republications. A trivial example may be the use of British English or American English. As discussed in Hoover 2019 these features made a significant difference in distinguishing the translators but can, at the same time, be very misleading, for example if the use of either British or American English is made by an editor afterwards or even be changed in a republication effort.
The issue of confounding variables has been discussed already in the earlier works (J. F. Burrows 2002; Hoover 2019) but more recently Caballero, Calvo, and Batyrshin 2021 tried to evaluate how strong confounding features influence methods of translator discrimination. They report results on two data sets where multiple translators translated the same original work. This way, the aforementioned issue is mitigated as we have samples with the following label assignment: (Author A, Translator B) and also (Author A, Translator C), for example. The first data set consists of three (complete) translations of the novel Don Quixote by Miguel de Cervantes from Spanish into English. The second data set consists of eight translations of plays by Henrik Ibsen from Norwegian into English. In this case, six plays were translated by only one translator and one play was translated twice, each time by a different translator.

trigram	**sample 1** ...
NOUN VERB ADJ	1 ...
VERB ADJ ,	1 ...
ADJ , AUX	1 ...
, AUX PROP	1 ...
AUX PROP.	1 ...
⋮	⋮ ⋱

Table 7.1: Example of the pos-punct-tri-gram bag of words as used in Caballero, Calvo, and Batyrshin 2021.

With these data sets, it is possible to train a model on one part, this one does not have to have multiple translations of the same original "non-parallel part", and evaluate it on the "parallel part" (i.e. multiple translations of the same original). Confounding variables that are specific to the original can then be ruled out, such as the genre or the topic or also highly specific features such as proper nouns that may otherwise be used by the model and inflate its predictive performance.
Caballero, Calvo, and Batyrshin 2021 found that their models work best on the parallel test set when the feature preprocessing included very strong regularization. In fact, the best-performing feature was what they called the "pos-punct-trigrams" which means for each text sample, tokens are replaced with their POS-tag (part of speech tag). This is done for all tokens except punctuations, – so the intermediate representation could look like this: 'NOUN VERB ADJ, AUX PROP. ...' Then, based on this intermediate representation, tri-grams are formed and put into a Bag-of-Words. The transformed example from above is depicted in Table 7.1. By replacing all words with their part of speech token, there is not much left of the text and anything that is related to the content or at least the topic of the text, is hidden. As stressed throughout this work, in DH usually there is not much data available, to discard so much of the source data in a preprocessing step is not a decision to be made lightly. Additionally, in Caballero, Calvo, and Batyrshin 2021 the data sets do not include a collaborative setting as it is the case for the Schlegel-Tieck translation. With an editor involved, who may normalize punctuations etc. post-translation the pos-punct-tri-gram features may not be as informative.

7.3 The Translation Corpus

The translation project involved 36 plays of which (to the best of our current knowledge) 17 were translated by A. Schlegel, 10 were translated by W. Baudissin, six were translated by D. Tieck. For the remaining three, D. Tieck and W. Baudissin collaborated.

> Im An=
> fang arbeitete ich mit Baudissin zusam=
> men, in Viel Lärmen um nichts sind die Ver=
> se von mir und die prosaischen Scenen von

> ihm. Die Widerspenstige haben wir beide
> ganz übersetzt, hernach ist von jedem das
> Beste behalten. Auf diese Art ging es
> aber zu langsam und machte sich auch nicht
> recht, weil wir eigentlich verschiedenen Grund=
> sätzen folgten, und wir theilten uns
> nun die Stücke. Ich bekam die Veroneser,
> Timon von Athen, Coriolan, Macbeth,
> Wintermährchen und Cymbeline. Coriolan
> und Macbeth haben mir die größte Freu=
> de gemacht. Baudissin hat viel Talent
> für das Leichte, Komische und die Wort=
> spiele, darum sind ihm auch die Irrungen und
> Love's labour's lost, was wir Liebes Lust und
> Leid genannt haben, vorzüglich gelungen, im
> letzteren sind einige Sonette von mir.[2]
> (BI, Dorothea Tieck to v. Uechtritz, Letter 8, pp. 8)

This quote from a letter by Dorothea Tieck gives very detailed insights into the collaborative setting and several observations need to be evaluated against in the following:

- The collaborative setting of the three plays in which they collaborated is different. Is it possible to reveal these differences with ML methods?

- D. Tieck focussed on the verse parts and W. Baudissin focussed on the prose parts. How much do the translators differ in style when only looking at each type? Also, along those lines: Is there a difference with regard to where the translators position themselves on the "Bertin axis" (estranging with the original or estranging with the reader) based on whether they translate prose or verse, specially when looking on what the translators think they are better at, considering their self-report.

- One thing that has to be handled with caution is that based on this quote, we can assume that they assigned the plays to each other such that the plays fit their translation style well. This may cause another confounder to be present, as we might reveal features of a play (that has therefore been assigned to a translator) instead of a feature of that particular translator themselves.

[2] In the beginning, I worked together with Baudissin, in Much Ado about Nothing I translated the verse and he translated the prosaic scenes. We both translated the Taming of the Shrew in its entirety and kept the best of each. It was too slow this way however and it also didn't really work out as we both had our principles, we then distributed the plays. I got the Two Gentlemen of Verona, Timon of Athens, Coriolanus, Macbeth, The Winter's Tale and Cymbeline. Coriolanus and Macbeth brought me the greatest joy. Because Baudissin has real talent for the light, comic and for word plays, the Comedy of Errors and Love's Labour's Lost, turned out beautifully. Love's Labour's Lost we translated with 'Liebes Lust und Leid' (Love's Desire and Suffering), in which I translated several sonnets.

As in the case of Love's Labour's Lost there are only a few very small parts written by Dorothea Tieck – some of the sonnets –, and since in this analysis scenes are used as the smallest unit of sample, this play was not put into the collaboration test set as any sample will be, by the most part, translated by W. Baudissin.

7.3.1 Bilingual Corpus

In the research on translationese it is common to analyse the original in comparison with the translation, as for example done in Baker, Francis, and Tognini-Bonelli 1993; Gellerstam 1986. In contrast, in the research on literary translator style as, for example Caballero, Calvo, and Batyrshin 2021; Rybicki and Heydel 2013 only the translation is considered although evidently it is a very obvious idea when analysing translator style to also take into account the original and compare what stylistic features have changed. One reason why this isn't done more regularly is that the methods for authorship attribution do not take two versions of the text as input. The methods need to be adapted to work with original and translation at the same time.

The second reason is that it is more complicated to design an NLP model that works in two languages. A method like the one proposed by Caballero, Calvo, and Batyrshin 2021 that uses pos-punct-tri-gram features (which means part of speech tags and punctuations) is very dependent on the language. In German, for example, the rules for commas are much more strict than in English. A translator may have more stylistic freedom when translating into English than when translating into German.

A third reason is that for the analysis of the difference between original and translation an aligned data set is needed. A data set that is aligned between the original and translation means that for a certain granularity (word, sentence, paragraph, chapter, etc.) it is annotated which unit from the original is translated into which unit from the translation. Evidently, creating these data sets is costly. Often, sentences are the chosen units for alignment. Trivially, in the act of translation not every sentence is translated as is but sometimes, for example, a sentence is split into two smaller sentences or two sentences are merged into one. It can also happen that sentences are discarded or new sentences are added, depending on how freely the translator translates. Sometimes aligned data sets are created manually but there are also automatic solutions available.[3]

However the translation alignment problem is more complicated: Original and translation cannot be aligned with an alignment algorithm that is used in computational biology because the original and the translation are not in the same language (and may not even use the same vocabulary). Therefore, an additional step has to be made prior to the actual alignment: Finding a unifying representation of the sentences (or other chosen level of granularity) of both languages. Sennrich 2011 propose a very simple approach to create a feature vector for each sentence: They utilize

[3]In fact, the problem is similar to the sequence alignment problem in biology where DNA or RNA sequences are aligned to find similar parts. In Computational Biology, one of the standard algorithms to find an alignment between two sequences is called the Needleman-Wunsch algorithm by Needleman and Wunsch 1970.

a machine translation model to translate the original into the target language (or vice versa). This way, the original and the translated sentences are present in the same language and can be compared with standard translation quality measure, such as the BLEU score by Papineni et al. 2002. Normally, the BLEU score is used to compare the prediction of a machine translation model with the ground truth translation (the higher the score, the more similar the automatic translation to the ground truth) so the hypothesis in this setting is that the better the machine learning model translates, the higher the bleu score between the automatic translation of a sentence and the sentence it should be aligned with. With this distance measure, a sequence alignment algorithm can be employed that imposes the typical constraints that are loaned from computational biology (such as sequentiality) but also have been used for a long time already in bilingual text alignment (Gale and Church 1993).
Joulin et al. 2018 propose a different approach to produce a common feature set for both languages: aligned word vectors. As described in Chapter 4, vectorial representations of words can be trained with unlabelled text that incorporate some syntactic and semantic information of the words. In Joulin et al. 2018, word vectors are not trained independently for different languages but they are trained such that word's translations have similar representations. This is done by using an initial dictionary of translations. With this method, words from original and translation can be embedded into the same vector space and standard distances such as eucledian or cosine distance can be computed that can serve as the distance measure similar to the BLEU score to be used in the sequence alignment algorithm.

7.4 Method

A total of three experiments were carried out, the first two dealing in particular with the question of the individual translation properties of D. Tieck and W. Baudissin, while the third experiment will assess the question of L. Tieck's contribution. The data layout and the analysis steps of all experiments are shown in Figure 2. The English corpus is retrieved from the *Digital facsimile of the Bodleian First Folio of Shakespeare's plays, Arch. G c.7* 2022, for the German corpus, TextGrid[4] was used. Throughout the experiments, the spacy tokenizer and lemmatizer was used (Montani et al. 2022). For counting the number of syllables per line, the pyphen[5] package was used. In the first experiment, solely on the basis of the German material, translation-stylistic characteristics are to be found that discriminate the translator. In addition to Nearest Neighbors on Burrows' Delta (J. Burrows 2002; Argamon 2008) that was used by Rybicki and Heydel 2013, Bag-of-N-Gram features and also pre-trained word vectors using the Fasttext model (Bojanowski et al. 2017) were used and classified by a Support Vector Machine with RBF kernels

[4]See `textgrid.de/`.
[5]See `https://pyphen.org/`.

(Cortes and Vapnik 1995; K.-R. Müller et al. 2001). Cross validation was used to find appropriate hyper parameters using scikit-learn (Pedregosa et al. 2011).
In the second experiment, we use the trained classifiers of Experiment 1 on the parts of the corpus where D. Tieck and W. Baudissin collaborated. We compute the predicted class of each scene individually and try to examine who the major translator of each part of the translation was. This explorative experiment enables us to concentrate on scenes for which the classifiers tend to agree, which we then manually evaluate.
In Experiment 3, cross-language features between the English material and its German translations are compared with respect to its translator. As shown in Figure 2, the first step for analysing the translation is to align the original and the translation, to be able to identify deviations on scene level. During the translation process, the scene boundaries were not always preserved and in order to compare intervals of the same contents, an automatic mapping of scenes is performed. Afterwards, three different features are compared:

1. *Richness* defines the ratio between types (unique tokens) and tokens in a sample. The larger the richness the more variety is present in the sample. A richness score = 1 means that no word occurs more than once in the sample.

2. *Syllables per line* defines the median number of syllables that occur in one line of text. This feature is especially important for the parts of the plays that are written (and translated) in verse form. It may reveal how a certain meter is translated.

3. *Burrows' delta* has already been discussed previously, it describes the distance of the normalized bag of words representation between two samples given a specific vocabulary. In this case, Burrows' delta is not computed across languages but it is computed pair-wise within a language and the difference between the deltas of the sample pairs are analyzed.

When explicitly modeling the translation as a textual variant in comparison to the original, the confounding variables that are specific to the original can be largely ruled out.

7.5 Results

Experiment 1: Classify translator of scenes in validation set: As shown in Table 1, the individual classifiers on scene level show decent performance. Burrows' Delta, however, does not show convincing results. For further improvement, we combined the classifiers by filtering scenes for which all scene-classifiers agree. This results in a smaller test set (57 scenes) but also in a considerable performance boost. For this subset of the test set, our combined classifier is on average performing with a precision and recall of ≈ 93 Overall, the classifiers perform better in identifying scenes by W. Baudissin.

Method		Burrows'	W-N-Grams	C-N-Grams	WV	Combined
Grouping		Play	Scene	Scene	Scene	Scene
D. Tieck	F1	50.00	62.16	64.86	79.52	89.47
	P	50.00	67.65	70.59	76.74	94.44
	R	50.00	57.50	60.00	82.50	85.00
	#	2	40	40	40	20
W. Baudissin	F1	66.67	77.05	78.69	84.96	94.74
	P	66.67	73.44	75.00	87.27	92.31
	R	66.67	81.03	82.76	82.76	97.30
	#	3	58	58	58	37
Weighted average	F1	60.00	70.97	73.05	82.74	92.89
	P	60.00	71.05	73.20	82.98	93.06
	R	60.00	71.43	73.47	82.65	92.98
	#	5	98	98	98	57

Table 7.2: Scores on held-out test set for various features and groupings. For classification of N-Gram features and Word Vectors, an SVM with RBF Kernel has been used. The Support row denotes the number of scenes in the respective class. Parameters have been optimized using grid search and 5-fold cross validation. For Burrows' delta, a Nearest Neighbors Classifier has been used. The optimal number of features for the delta has been cross validated as well. The best method is highlighted in grey for the three competing methods on scene level. F1, precision and recall are reported by a factor of 100.

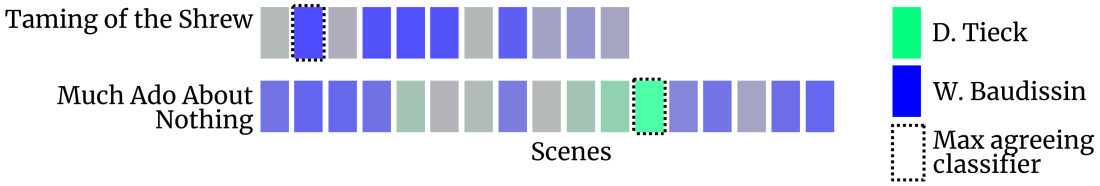

Figure 7.1: Average score of all scene-level classifiers of Experiment 1 to attribute each scene to D. Tieck or W. Baudissin for the two plays in which they collaborated.

Experiment 2: Classify translator of scenes in the collaboration set: In Figure 7.1, the translator attribution for the collaborative scenes are shown. Additionally, we exploit the finding of Experiment 1 that our classifiers performance is boosted when they are combined. In Viel Lärmen um nichts (Much adoe about Nothing), fourth act, first scene the highest agreement for D. Tieck, in Der Widerspenstigen Zähmung (The Taming of the Shrew) first act, second scene the highest agreement for Baudissin is observed. As it turns out, the two scenes are exceptionally long scenes with 302 and 264 speeches respectively, although the mean number of speeches per scene over the whole German corpus is only ≈118.7. The length of the scene may give the classifiers more features to distinguish the translators. The scene from The Taming of the Shrew alternates between Verses and Prose which may have given the translator the chance to underline their characteristic style. The scene from Much adoe about Nothing has a much more coherent rhythm which possibly fits D. Tieck's translation style better.

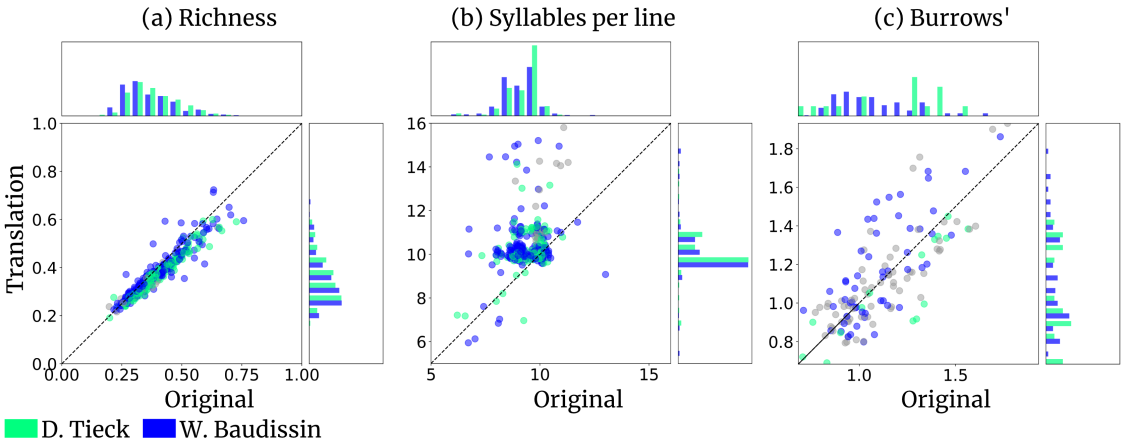

Figure 7.2: Three different features that compare original texts and their translations across languages. For each panel, the horizontal axis corresponds to the original version in English, the vertical axis corresponds to the German translation. The richness feature (a) shows little deviation in both languages. The Syllables per line feature (b) shows deviation in the translation for both translators and the Burrows' feature (c) shows deviation especially for one translator: D. Tieck (Green). For (b) gaussian noise (with std. of .2) was added to the points to visualize overlapping points. Also, in (b), a few outliers are not visualized. The points in (c) are grey if both plays were not translated by the same person.

Experiment 3: Identify Contribution of Ludwig Tieck: In Figure 7.2, the results of the cross-language comparison are shown. Points in all panels that are close to the diagonal do not deviate across language. The richness (a) of the scenes stay very close to the diagonal, however the majority of points is slightly below the diagonal. The original is slightly "richer" in the sense of our measurement than the translation, but there is no difference across translators. The median syllables per line (b) of the translation deviates quite significantly in that the German version often uses more syllables per line than the English version. D. Tieck stated in her letter that she also translated Sonnets even in a play that was otherwise translated by W. Baudissin. Because of this statement we originally expected D. Tieck to follow the number of syllables of the original more strictly. This expectation is also in line with the findings of Experiment 2 where most classifiers agree on D. Tieck as the translator in a scene with a coherent rhythm. However, the findings of (b) cannot verify this hypothesis, because the deviation exists for both translators. In (c), the points visualize Burrows' delta between the two plays in English, the vertical position is the Burrows' delta of the respective pair in German. Each data point for which both plays are translated by the same person is color-coded accordingly (grey otherwise). Interestingly, the green points are almost exclusively below the diagonal, with only a few exceptions for plays that already exhibit a small delta in the English version. This indicates translations by D. Tieck move closer to each other and thus may incorporate a more consistent style.

7.6 Summary & Conclusion

We proposed an ensemble of translator attribution methods that result in a very high performance on scenes where they agree (Experiment 1). We show a significant improvement over state of the art methods for translator attribution. This combination of classifiers is used to suggest translators for scenes where the true translator is unknown. A close reading of the scenes revealed distinct characteristics that could explain the decision of the classifiers (Experiment 2). We thus argue that this method likely found scenes where the majority of translation work can be attributed to the proposed translator. When looking at the results of the individual classifiers of Experiment 1, it is evident that the performance varies a lot. Only when focussing on the samples where the classifiers agree, the performance increases. Future work should be aimed toward finding more definitive predictions for the samples where the classifiers currently disagree.

A novel approach of comparing the material in the source language and the translations yield the result that D. Tieck has a more distinct style in her translations (Experiment 3, c). This showcases that the comparison of textual variants is a useful addition to the methodology of translator style attribution as it helps with eliminating confounders. With regard to the daughter–father relationship this can be seen as a literary independence from her father. Also, it could be observed that there is a translation system on which the three collaborators agree (Experiment 3, a and b). In that, we identified candidate features that could signal a contribution of L. Tieck. For further analysis we plan to include original plays by L. Tieck in order to identify distinct characteristics that further narrow down his contribution to the translation. We also plan to include additional cross-language features that characterize a distinct style of W. Baudissin.

For Experiment 3, unfortunately only some combinations of deviations from the diagonal give insights: we can, for example, sometimes not know if a deviation from the diagonal was a decision made by the editor or if it resulted from a congeniality of the translators. This is a logical limitation that, considering this data set, there is not more information available. However, it might be possible to include additional data, for example authorial documents written by the editor to compare their style. This way, it might be possible to unravel the collaborative setting in even greater detail.
Another experiment that was originally planned to be included was to compare the style of D. Tiecks letters with the style of her translations. In this case, the style of an author would be compared with the style of a translator. This would result in to actually finding a style of a person that can write in different functions (translator and author). Besides the difficulty to find a style of a person regardless of whether translating or writing as an author, the main road block was the stark difference in document type (translation of a Shakespeare play vs. letter to a friend).

Overall, this chapter shows how approaches to the specific research question of analyzing translator style are shifted by embedding them in the methodology of textual variants, giving them a new perspective. While these tie in with established methods such as Burrows' Delta, which has also been used in translator attribution, it uses them in novel ways on aligned bilingual corpora so that comparability across language can be established.

Chapter 8

Summary and Conclusion

The goal of this thesis was to investigate complex digital humanities corpora and their documents with numerous types of textual variants using machine learning methods. Linear representations of text that are common in NLP are a reasonable simplification in many cases but especially for applications in the digital humanities, such as the analysis of genetic editions or translation variants, more complex representations are necessary. There are three main aspects that have been addressed in this thesis, improving the availability of digital scholarly editions and their interoperability with NLP tools (Part 1), the methodological alignment between NLP and literary studies for the concept of textual representations and the proposition of novel textual representation methods for structured corpora (Part 2) and the development of machine learning methods that are capable of processing complex document structures (Part 3). In this chapter, the main contributions of the thesis are summarized and it is concluded what the limitations are and how this can be leveraged for future work.

Part 1: In order to improve interoperability between digital scholarly editions and NLP tools, in Chapter 2 the Standoff Converter was proposed that enables scholars to programmatically extract a research-specific, linear textual variants from non-linear documents of digital scholarly editions.
Regarding the availability of datasets in the DH, three problems were identified. First, the creation of new datasets is expensive because it requires human labor. Second, for many digital sources that are eligible for use in editions, the legal situation regarding copyright is unclear, and third, many digital editions use their own standards that are not consistent with each other. This makes it difficult to combine or integrate corpora.
In Chapter 3, we have therefore explored the extent to which machine learning models can assist humans in creating and annotating datasets to make the process more efficient. We have also clarified the legal situation for the reuse and republication of many digital sources not previously used in editions, and third, we have outlined a working path for publishing these sources in a standardized and principled way to facilitate their use in other research projects or to make it at all possible.
Outlook: Even though important steps have been taken in this direction, it has to be

stated that there is and will continue to be a substantial lack of high quality digital scholarly editions. In order to increase the availability of datasets and the use of machine learning methods, we believe that future progress must be made in three directions in particular: There is a need to continue to invest work in the creation of new digital editions, as well as to continue to support initiatives working on conversion tools, such as the Standoff Converter, and standardizations for interoperability between different editions. Third, realistically, however, the status quo will continue to be that there are significant gaps in the available corpora, and despite the other two directions of advancement, these gaps will not be resolved in the foreseeable future. Therefore, it is important that the consideration of the incompleteness of the data becomes even more a part of the methodology of the digital humanities toward increasing robustness of the methodologies, (See Gengnagel 2022).

Part 2: In this part, we addressed the main obstacle of different methodologies in machine learning and literary studies. Therefore both views need to be take into account in order to reach both fields. In order to unify the view on the methodologies of the two disciplines the concept of textual representations was identified to be a central hinge point.

We introduced the fundamentals of text representations from both viewpoints and discussed how textual representations can be used in the context of 'machine learning reading' in Chapter 4, we then identified that a major shortcoming of existing word embedding approaches is that structural aspects between parts of the subject corpus are not adequately modeled, Chapter 5 therefore introduces the novel Word2Vec with Structure Prediction method that models both word representations and corpus structure, simultaneously.

This enabled us to create textual representations that are specific to sub corpora, in our case, famous German authors from the nineteenth century and to discover similarities between them on a macro-scale. The same representations could also be used to extract individual word embeddings for arbitrary downstream tasks.

Outlook: Unfortunately, these methods, even if popular in the sub-field of computational literary studies, are not yet widespread across the traditional literary studies and there are even distinct voices that question the usefulness of computational methods in the literary studies (See Da 2019).

When it comes to the methods for text representation that were proposed, a major shortcoming, is that they only have a very limited form of operationalising the intended reading compared to a literary critic. Selecting an appropriate reference corpus as a reading horizon is not as specific as, for example, 'reading through the lens of Marxism', as selected corpora might contain various aspects in addition to the intended one.

Likewise, when the size of the reference corpora increase, this bears the risk of unwanted training data points and harmful representations. For the machine learning community, the focus on curating and documenting the data sets presents an opportunity to address ethical questions which have been neglected for the most part.

The first attempts have already been made by explicitly asking library scholars to help with the data curation process also for purely ML or NLP research (Jo and Gebru 2020).
In the archival sciences, it has also been criticized lately that colonial practices have harmed (and are harming) archiving in numerous communities around the world which lead to silencing of those communities' cultural voices (Levi 2022). So although using ML in the context of the computational literary studies has the potential of reducing biases, it is just that: a potential that has to be realised.
When it comes to Word2Vec with Structure Prediction, it was identified that it can be sensitive to sub-corpora with fewer data and might picture them as outliers in the predicted structure. It would be worthwhile to study how to better control the likely trade-off between structure embedding and word embedding quality specifically for outliers.

Part 3: In this part, it was identified that even if sufficient amounts of textual variants extracted from digital scholarly editions exist and the research question was methodologically aligned between DH and ML, there was still a lack of specific machine learning methods that are able to jointly analyse the different textual variants.
With the analysis of both the Berlin Intellectuals edition and the famous Schlegel-Tiecksche Shakespeare translations, two research projects were conducted that showed the importance of comparing textual variants and the importance of customized and robust machine learning methods in the context of DH work. When analysing the letters from the Berlin Intellectuals edition with the novel AlterLDA method, previously suspected alteration reasons could be evaluated and largely confirmed.
Additionally, when analysing the original plays by Shakespeare in comparison with their translations by different translators certain features could be identified to be characteristic of one translator, Dorothea Tieck who, to our current knowledge, did not publish authorial works of her own. This is especially interesting as it suggests that she developed her personal translation style without having an author style.
Outlook: The main aspect that hinders the widespread use of the AlterLDA method as of now is that there are only few data sets that are so diligently annotated to be able to apply it. This echoes what we have vocalised previously, that an initiative to advertise the adoption of annotations schemas for document structure in both communities NLP and DH would not only incentivize the NLP community to develop more tools that take into account these complex structures thereby helping to curate training corpora for NLP but it would also contribute to a better understanding of literary text genesis in other contexts.
More specific to the Berlin Intellectuals edition, it has been shown that the multilinguality of the corpus had unexpected effects on the results of the AlterLDA model. It would be a fruitful addition to see new approaches emerge that take into account different languages more explicitly.

Most of the translator style research still considers only the translated text and does not compare the translations to the original. Future research should be directed toward a fine-grained alignment of the source and the target of the translation to be able to conduct the stylistic analyses that were conducted in Chapter 7 also on other sources.
With regard to classification of the translation only, for some of the samples the method could not give any robust answer yet. Evidently, the stronger the confounding variables are, such as comparing writing in prose and writing in verse, the more the performance of the methods decrease. This is expected as a translation is generally not independent of these factors.
This leads, however, to the fascinating question of writing style: How exactly do author style and translator style relate? Especially in the case where authors also translated the question arises, what constitutes a personal writing style.
From a modeling perspective, the most relevant question is how we can incorporate more complex features than simple counts of punctuations and alike into the model without overfitting on confounding variables? We believe that supervised pre-training and unlearning are promising paths forward.

All in all, when considering that one of the shortcomings of the current AlterLDA model is that it sometimes fails with multi-linguality and also seeing that the research on alterations of manuscripts and the research on translator styles are both analyzed with the same framework (textual transformations), a promising future path of research would be to unify both aspects. This unifying view would allow to have holistic perspective on the history of textual documents and would be able to attribute textual phenomena to all actors that contributed to the subject variant, let it be editors, translators or authors.

Glossary

canon is a corpus of literary texts that is considered valuable and worth preserving (Vgl. V. Nünning and A. Nünning 2010, p. 31). 3, 17

ego document s are texts written in the first person that document personal experience in their historical context. They include letters, diaries, memoirs and have gained momentum as a primary source in historical research and literary studies over the past decades. 66–68, 71

genetic edition is a type of scholarly edition that focusses on how the text was created, not on how the text currently is or how it was intended to be. It therefore focusses on the material, such as the description of manuscripts to unwind the history of changes that were made. 6, 72, 73, 101

plain text is a sequence of characters or symbols, for example a variant the text extracted from a TEI document that can be used as input for an NLP pipeline. 3, 6, 8, 10, 11, 13, 15, 16, 36, 38, 124, 125

radical transformation is a term used by Ramsay 2011 that describes transformations performed on a text by a reading of the text. 43, 46, 74

reference corpus is a collection of documents that is used as supplementary data for training representations or pre-training. It is meant to be selected in a specific relation to the subject corpus. 39, 45–47, 49, 102

subject corpus is the collection of documents that is selected to be immediately relevant to the research question. 39, 45, 65, 102

translationese is a term coined by Gellerstam 1986 that describes the characteristics of texts that were translated into a specific language in comparison to texts originally written that language. 90, 94

type s are the elements in the set of unique tokens of a corpus - the vocabulary. 40, 41, 44, 96

Acronyms

AlterLDA Alteration Latent Dirichlet Allocation. 3, 74–82, 84–87, 103, 104, 136, 137

ALTO analyzed layout and text object. 16

BI Berlin Intellectuals (Baillot 2022). 10, 72–75, 79, 80, 83–87, 93, 138, 140

BPE byte-pair encoding. 38

CHI cultural heritage institution. 18, 20, 21, 23, 31

CLS computational literary studies. 44, 47, 49, 103

DH digital humanities. 1–3, 7, 15, 18, 31, 68, 74, 92, 103

HTR handwritten rext recognition. 20

LDA Latent Dirichlet Allocation. 3, 77, 78, 80

LLM large language model. 18, 19

METS metadata encoding & transmission standard. 24–28

ML machine learning. 2, 3, 17, 19, 34, 47, 93, 103

NER named entity recognition. 10, 18, 39

NER/L named entity recognition and linking. 18

NLP natural language processing. 1, 2, 4, 6, 10–16, 18, 19, 46, 68, 94, 103

OCR optical character recognition. 2, 8, 16–21, 23, 24, 28–31, 35, 128

PAGE page analysis and ground-truth elements. 16, 26–30

SVG scalable vector graphics. 16

TEI text encoding initiative. 4, 7, 8, 10–16, 73, 74

XML extensible markup language. 7, 11, 14, 27, 28, 30, 72, 74, 124

Bibliography

Akbik, Alan, Duncan Blythe, and Roland Vollgraf (2018). "Contextual String Embeddings for Sequence Labeling". In: *Proceedings of the International Conference on Computational Linguistics*. 27. Association for Computational Linguistics, pp. 1638–1649.

Andrews, Tara (2013). "The third way: philology and critical edition in the digital age". In: *Variants* 10, pp. 61–76.

Argamon, Shlomo (2008). "Interpreting Burrows's Delta: Geometric and probabilistic foundations". In: *Literary and Linguistic Computing* 23.2, pp. 131–147.

Ast, Friedrich (1808). *Grundlinien der Grammatik, Hermeneutik und Kritik*. Landshut: Jos. Thomann'sche Buchdruckerei.

Azarbonyad, Hosein, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps (2017). "Words are malleable: Computing semantic shifts in political and media discourse". In: *Proceedings of the Conference on Information and Knowledge Management*, pp. 1509–1518.

Bachleitner, Norbert (1989). "'Übersetzungsfabriken'. Das deutsche Übersetzungswesen in der ersten Hälfte des 19. Jahrhunderts". In: *Internationales Archiv für Sozialgeschichte der deutschen Literatur*, pp. 1–50.

Baillot, Anne (2016). "Berliner 'Intellektuelle' um 1800. Eine kontroverse Kategorie und ihre Anwendbarkeit im digitalen Zeitalter". In: *Virtuosen der Öffentlichkeit? Friedrich von Gentz (1764-1832) im globalen intellektuellen Kontext seiner Zeit*. Ed. by Gudrun Gersmann, Friedrich Jaeger, and Michael Rohrschneider. mapublishing.

– (2018). "Die Krux mit dem Netz Verknüpfung und Visualisierung bei digitalen Briefeditionen". In: *Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven*. Ed. by Toni Bernhart, Marcus Willand, Sandra Richter, and Andrea Albrecht. De Gruyter, p. 355–370.

– ed. (2022). *Letters and texts. Intellectual Berlin around 1800*. URL: `berliner-intellektuelle.eu` (visited on 11/01/2022).

Baillot, Anne and Anna Busch (2014). "'Berliner Intellektuelle um 1800' als Programm. Über Potential und Grenzen digitalen Edierens". In: *literaturkritik* 9.

– (2015). "Editing for Man and Machine: The Digital Edition Letters and Texts. Intellectual Berlin around 1800 as an Example". In: *Variants* 15-16.

Baillot, Anne and David Lassner (2022). "Von Graphen zu Word Embeddings. Zur Entwicklung des mathematischen und visuellen Instrumentariums der Literaturwissenschaft". In: *Germanica* 71 (2), pp. 191–203.

Baillot, Anne, Mike Mertens, and Laurent Romary (2016). "Data fluidity in DARIAH–pushing the agenda forward". In: *BIBLIOTHEK Forschung und Praxis* 40.2, pp. 151–165.

Baillot, Anne and Markus Schnöpf (2015). "Von wissenschaftlichen Editionen als interoperable Projekte, oder: Was können eigentlich digitale Editionen?" In: *Historische Mitteilungen der Ranke-Gesellschat*. Die Zukunft der Digital Humanities Beiheft 91, pp. 139–156.

Baker, Mona, Gill Francis, and Elena Tognini-Bonelli (1993). *Text and technology: in honour of John Sinclair*. John Benjamins Publishing.

Bamler, Robert and Stephan Mandt (2017). *Dynamic word embeddings*. arXiv preprint, 1702.08359.

Beckett, Samuel (2022). *Digital Manuscript Project. A digital Genetic Edition*. Ed. by James Little, Vincent Neyt, Shane Van Hulle Dirk ans Weller, Pim Verhulst, Mark Nixon, and Magessa O'Reilly. University Press Antwerp. URL: `http://www.beckettarchive.org`.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜". In: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623.

Bertin, Jacques (1983). "Semiology of graphics: Diagrams, networks, maps". Trans. by W. J. Berg. In: *The University of Wisconsin Press, Ltd.*

Binder, Alexander, Klaus-Robert Müller, and Motoaki Kawanabe (2012). "On Taxonomies for Multi-class Image Categorization". In: *International Journal of Computer Vision* (3), pp. 281–301.

Birhane, Abeba and Vinay Uday Prabhu (2021). "Large Image Datasets: A Pyrrhic Win for Computer Vision?" In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1537–1547.

Blei, David, Andrew Ng, and Michael Jordan (2003). "Latent Dirichlet allocation". In: *Journal of Machine Learning Research* 3, pp. 993–1022.

Blei, David M., Thomas L. Griffiths, and Michael I. Jordan (2010). "The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies". In: *Journal of the Association for Computing Machinery* 57.2, pp. 1–30.

Bleich, Erik, Hasher Nisar, and Rana Abdelhamid (2016). "The effect of terrorist events on media portrayals of Islam and Muslims: evidence from New York Times headlines, 1985–2013". In: *Ethnic and Racial Studies* 39.7, pp. 1109–1127.

Boenig, Matthias, Konstantin Baierer, Volker Hartmann, Maria Federbusch, and Clemens Neudecker (2019). "Labelling OCR Ground Truth for Usage in Repositories". In: *Proceedings of the International Conference on Digital Access to Textual Cultural Heritage*. 3, pp. 3–8.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017). "Enriching word vectors with subword information". In: *Transactions of the association for computational linguistics* 5, pp. 135–146.

Bommasani, Rishi et al. (2021). *On the Opportunities and Risks of Foundation Models*. arXiv preprint, 2108.07258.

Burrows, John (2002). "'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship". In: *Literary and Linguistic Computing* 17.3, pp. 267–287.

Burrows, John F. (2002). "The Englishing of Juvenal: Computational stylistics and translated texts". In: *Style* 4.36, pp. 677–750.

Caballero, Christian, Hiram Calvo, and Ildar Batyrshin (2021). "On explainable features for translatorship attribution: Unveiling the translator's style with causality". In: *IEEE Access* 9, pp. 93195–93208.

Calvo Tello, José (2021). *The Novel in the Spanish Silver Age*. Bielefeld University Press.

Carpenter, Bob (2010). *Integrating out multinomial parameters in latent Dirichlet allocation and naive Bayes for collapsed Gibbs sampling*. Tech. rep. LingPipe.

Chagué, Alix and Thibault Clérice (2022). *HTR-United*. URL: `https://htr-united.github.io/` (visited on 04/19/2022).

Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks". In: *Machine learning* 20.3, pp. 273–297.

Cultural Heritage Data Reuse Charter (2022). DARIAH-EU. URL: `https://www.dariah.eu/activities/open-science/data-re-use/` (visited on 04/19/2022).

Da, Nan Z. (2019). "The Computational Case against Computational Literary Studies". In: *Critical Inquiry* 3, pp. 601–639.

Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman (1990). "Indexing by latent semantic analysis". In: *Journal of the American society for information science* 41.6, pp. 391–407.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol. 1, pp. 4171–4186.

Digital facsimile of the Bodleian First Folio of Shakespeare's plays, Arch. G c.7 (2022). URL: `https://firstfolio.bodleian.ox.ac.uk/`.

Dobson, James E (2015). "Can An Algorithm Be Disturbed?: Machine Learning, Intrinsic Criticism, and the Digital Humanities". In: *College Literature* 42.4, pp. 543–564.

– (2021). "Vector hermeneutics: On the interpretation of vector space models of text". In: *Digital Scholarship in the Humanities* 37.1, pp. 81–93.

Drucker, Johanna (2017). "Non-representational approaches to modeling interpretation in a graphical environment". In: *Digital Scholarship in the Humanities* 33.2, pp. 248–263.

Eberle, Oliver, Jochen Büttner, Florian Kräutli, Klaus-Robert Müller, Matteo Valleriani, and Grégoire Montavon (2022). "Building and Interpreting Deep Similarity Models". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.3, pp. 1149–1161.

Eclevia, Marian Ramos, John Christopher La Torre Fredeluces, Carlos jr Lagrosas Eclevia, and Roselle Saguibo Maestro (2019). "What Makes a Data Librarian?" In: *Qualitative and Quantitative Methods in Libraries* 8.3, pp. 273–290.

Ehrmann, Daniel (2016). "Textrevision–Werkrevision. Produktion und Überarbeitung im Wechsel von Autoren, Herausgebern und Schreibern". In: *Editio* 30.1, pp. 71–87.

Engl, Elisabeth, Konstantin Baierer, Matthias Boenig, Volker Hartmann, and Clemens Neudecker (2020). "Volltexte – die Zukunft alter Drucke: Bericht zum Abschlussworkshop des OCR-D-Projekts". In: *o-bib. Das offene Bibliotheksjournal* 7.2, pp. 1–4.

Faruqui, Manaal and Chris Dyer (2014). "Community evaluation and exchange of word vectors at wordvectors.org". In: *Proceedings of the Association for Computational Linguistics: System Demonstrations*. 52, pp. 19–24.

Fellbaum, Christiane, ed. (1998). *WordNet: An electronic lexical database*. Cambridge: MIT Press.

Flaubert, Gustave (2009). *Les Manuscrits de Madame Bovary: Édition intégrale sur le web*. Ed. by Danielle Girard and Yvan Leclerc. Rouen: Université de Rouen.

Freytag, Gustav (1863). *Die Technik des Dramas*. Leipzig: Verlag von S. Hirzel.

Fuchs, Yvonne and Dominic Weber (2022). *Transcriptiones. A platform for hosting, accessing and sharing transcripts of non-digitised historical manuscripts.* ETH-Library. URL: https://www.librarylab.ethz.ch/de/project/transcriptiones/ (visited on 04/19/2022).

Gadamer, Hans-Georg (1975). *Hermeneutik I. Wahrheit und Methode: Grundzüge einer philosophischen Hermeneutik*. Verlag Mohr Siebeck.

Gage, Philip (1994). "A New Algorithm for Data Compression". In: *C Users Journal*, pp. 23–38.

Gale, William A. and Kenneth W. Church (1993). "A Program for Aligning Sentences in Bilingual Corpora". In: *Computational Linguistics* 19.1, pp. 75–102.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou (2018). "Word embeddings quantify 100 years of gender and ethnic stereotypes". In: *Proceedings of the National Academy of Sciences* 115.16, E3635–E3644.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford (2021). "Datasheets for Datasets". In: *Communications of the Association for Computing Machinery* 64.12, pp. 86–92.

Gellerstam, Martin (1986). "Translationese in Swedish novels translated from English". In: *Translation studies in Scandinavia* 1, pp. 88–95.

Gengnagel, Tessa (2022). *Digital Humanities, or: The Broken Record of Everything.* URL: http://web.archive.org/web/20220821214305/https://www.youtube.com/watch?v=G3Nn8gw81cA.

Goethe, Johann Wolfgang von (2022). *Faust. Historisch-kritische Edition*. Ed. by Anne Bohnenkamp, Silke Henke, and Fotis Jannidis. Frankfurt am Main / Weimar / Würzburg. Version 1.2rc.

Gonen, Hila, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg (2020). "Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 58, pp. 538–555.

Grave, Edouard, Armand Joulin, and Quentin Berthet (2019). "Unsupervised alignment of embeddings with wasserstein procrustes". In: *Proceedings of the International Conference on Artificial Intelligence and Statistics*. 22, pp. 1880–1890.

Gruber, Amit, Yair Weiss, and Michal Rosen-Zvi (2007). "Hidden Topic Markov Models". In: *Proceedings of the International Conference on Artificial Intelligence and Statistics*. Vol. 2. 11, pp. 163–170.

El-Hajj, Hassan et al. (2022). "An Ever-Expanding Humanities Knowledge Graph: The Sphaera Corpus at the Intersection of Humanities, Data Management, and Machine Learning". In: *Datenbank-Spektrum*, pp. 153–162.

Hamdi, Ahmed, Axel Jean-Caurant, Nicolas Sidère, Mickaël Coustaty, and Antoine Doucet (2020). "Assessing and minimizing the impact of OCR quality on named entity recognition". In: *International Conference on Theory and Practice of Digital Libraries*, pp. 87–101.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky (2016). "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change". In: *Proceedings of the Association for Computational Linguistics*. 54, pp. 1489–1501.

Heuser, Ryan (2020). *Abstraction: A Literary History*. URL: `https://ryanheuser.org/talks/kingscollege2020/` (visited on 11/01/2022).

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780.

Hofmann, Valentin, Janet B Pierrehumbert, and Hinrich Schütze (2020). *Dynamic Contextualized Word Embeddings*. arXiv preprint, 2010.12684.

Hoover, David L. (2019). "The Invisible Translator Revisited". In: *Book of Abstracts of the Digital Humanities Conference*.

Hutchinson, Ben, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl (2020). "Social Biases in NLP Models as Barriers for Persons with Disabilities". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 58, pp. 5491–5501.

Jannidis, Fotis and Gerhard Lauer (2014). "Burrows's Delta and Its Use in German Literary History". In: *Distant Readings. Topologies of German Culture in the Long Nineteenth Century*. Ed. by Matt Erlin and Tatlock Lynne. Camden House, pp. 29–54.

Jawahar, Ganesh and Djamé Seddah (2019). "Contextualized Diachronic Word Representations". In: *Proceedings of the International Workshop on Computational Approaches to Historical Language Change*. 1, pp. 35–47.

Jelodar, Hamed, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao (2019). "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey". In: *Multimedia Tools and Applications* 78.11, pp. 15169–15211.

Jo, Eun Seo and Timnit Gebru (2020). "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 306–316.

Joulin, Armand, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave (2018). "Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 2979–2984.

Kiessling, Benjamin (2019). "Kraken – an Universal Text Recognizer for the Humanities". In: *Book of Abstracts of the Digital Humanities Conference*.

Kiessling, Benjamin, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra (2019). "eScriptorium: An Open Source Platform for Historical Document Analysis". In: *Proceedings of the International Conference on Document Analysis and Recognition Workshops*. Vol. 2, pp. 19–19.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov (2014). *Temporal analysis of language through neural language models*. arXiv preprint, 1405.3515.

Kingma, Diederik P and Jimmy Ba (2014). *Adam: A method for stochastic optimization*. arXiv preprint, 1412.6980.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena (2015). "Statistically significant detection of linguistic change". In: *Proceedings of the International Conference on World Wide Web*. 24, pp. 625–635.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal (2018). "Diachronic word embeddings and semantic shifts: a survey". In: *Proceedings of the International Conference on Computational Linguistics*. 27, pp. 1384–1397.

Lansdall-Welfare, Thomas, Saatviga Sudhahar, James Thompson, Justin Lewis, FindMyPast Newspaper Team, and Nello Cristianini (2017). "Content analysis of 150 years of British periodicals". In: *Proceedings of the National Academy of Sciences* 114.4, E457–E465.

Lassner, David, Anne Baillot, and Julius Coburger (2019). "Attributions Of Early German Shakespeare Translations". In: *Book of Abstracts of the Digital Humanities Conference*. DOI: 10.34894/DK6QKN.

Lassner, David, Anne Baillot, Sergej Dogadov, Klaus-Robert Müller, and Shinichi Nakajima (2021). "Automatic Identification of Types of Alterations in Historical Manuscripts". In: *Digital Humanities Quarterly* 15.2, online. URL: http://www.digitalhumanities.org/dhq/vol/15/2/000553/000553.html.

Lassner, David, Stephanie Brandl, Anne Baillot, and Shinichi Nakajima (2023). "Domain-Specific Word Embeddings with Structure Prediction". In: *Transactions of the Association for Computational Linguistics* 11, pp. 320–335.

Lassner, David, Julius Coburger, Clemens Neudecker, and Anne Baillot (2021). "Publishing an OCR ground truth data set for reuse in an unclear copyright setting". In: *Zeitschrift für digitale Geisteswissenschaften* Sonderband 5, Fabrikation von Erkenntnis – Experimente in den Digital Humanities, online. DOI: 10.17175/sb005_006.

– (2022). *Data set of the paper "Publishing an OCR ground truth data set for reuse in an unclear copyright setting"*. Version 1.1. URL: `https://zenodo.org/record/4742068` (visited on 04/19/2022).

Levi, Amalia S. (2022). "Review: Archival Silences: Missing, Lost and, Uncreated Archives". In: *The Public Historian* 44.1. Ed. by Michael Moss and David Thomas, pp. 113–116.

Levy, Omer and Yoav Goldberg (2014). "Neural word embedding as implicit matrix factorization". In: *Advances in neural information processing systems*. 27, pp. 2177–2185.

Li, Xiaoya, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li (2020). "Dice Loss for Data-imbalanced NLP Tasks". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 58, pp. 465–476.

Liebl, Bernhard and Manuel Burghardt (2020). "From Historical Newspapers to Machine-Readable Data: The Origami OCR Pipeline". In: *Computational Humanities Research*, pp. 351–373.

Liu, Lin, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou (2016). "An overview of topic modeling and its current applications in bioinformatics". In: *SpringerPlus* 5.1608, online.

Liwicki, Marcus, Alex Graves, Horst Bunke, and Jürgen Schmidhuber (2007). "A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks". In: *Proceedings of the International Conference on Document Analysis and Recognition*. 9, pp. 367–371.

Marjanen, Jani, Lidia Pivovarova, Elaine Zosa, and Jussi Kurunmäki (2019). "Clustering ideological terms in historical newspaper data with diachronic word embeddings". In: *International Workshop on Computational History, HistoInformatics*. 5, online.

Meirelles, Isabelle (2019). "Visualizing information". In: *The Shape of Data in the Digital Humanities. Modeling Texts and Text-based Resources*. Ed. by Julia Flanders and Fotis Jannidis. Routledge, pp. 167–177.

METS. Metadata Encoding & Transmission Standard (2022). The Library of Congress. URL: `http://www.loc.gov/standards/mets/` (visited on 04/19/2022).

Michel, Jean-Baptiste et al. (2011). "Quantitative Analysis of Culture Using Millions of Digitized Books". In: *Science* 331.6014, pp. 176–182.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv preprint, 1301.3781.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013). "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*, pp. 3111–3119.

Montani, Ines et al. (2022). *explosion/spaCy*. Version v3.4.1. URL: `https://zenodo.org/record/6907665`.

Montavon, Grégoire, Wojciech Samek, and Klaus-Robert Müller (2018). "Methods for interpreting and understanding deep neural networks". In: *Digital Signal Processing* 73, pp. 1–15.

Moretti, Franco (2005). *Graphs, maps, trees: abstract models for a literary history*. London and New York: Verso.

Müller, Klaus-Robert, Sebastian Mika, Gunnar Ratsch, Koji Tsuda, and Bernhard Schölkopf (2001). "An introduction to kernel-based learning algorithms". In: *Transactions on Neural Networks* 12.2, pp. 181–201.

Nakajima, Shinichi, Kazuho Watanabe, and Masashi Sugiyama (2019). *Variational Bayesian Learning Theory*. Cambridge University Press.

Needleman, Saul B. and Christian D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". In: *Journal of Molecular Biology* 48.3, pp. 443–453.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman (2020). "Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection". In: *Proceedings of the Language Resources and Evaluation Conference*. 12, pp. 4034–4043.

Nünning, Vera and Ansgar Nünning (2010). *Methoden der literatur- und kulturwissenschaftlichen Textanalyse*. J. B. Metzler.

Padilla, Thomas, Laurie Allen, Hannah Frost, Sarah Potvin, Russey Roke, Elizabeth, and Stewart Varner (2019). *Final Report – Always Already Computational: Collections as Data*. URL: `https://zenodo.org/record/3152935` (visited on 04/19/2022).

Paisley, John, Chong Wang, David M. Blei, and Michael I. Jordan (2015). "Nested Hierarchical Dirichlet Processes". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.2, pp. 256–270.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). "BLEU: A Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the Association for Computational Linguistics*. 40, pp. 311–318.

Paulin, Roger (1998). "Luise Gottsched und Dorothea Tieck. Vom Schicksal zweier Übersetzerinnen". In: *Shakespeare Jahrbuch* 134. Ed. by Wolfgang Weiss, pp. 108–122.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning (2014). "Glove: Global vectors for word representation". In: *Proceedings of the Conference on Empirical ethods in Natural Language Processing*, pp. 1532–1543.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). "Deep Contextualized Word Representations". In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2227–2237.

Phuong, Mary and Marcus Hutter (2022). *Formal Algorithms for Transformers*. arXiv preprint, 2207.09238.

Plachta, Bodo (2006). *Editionswissenschaft: eine Einführung in Methode und Praxis der Edition neuerer Texte*. 2nd ed. Stuttgart: Reclam.

Pletschacher, S. and A. Antonacopoulos (2010). "The PAGE (Page Analysis and Ground-Truth Elements) Format Framework". In: *Proceedings of the International Conference on Pattern Recognition*. 20, pp. 257–260.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever (2018). *Improving language understanding with unsupervised learning*. Tech. rep. OpenAI.

Ralle, Inga Hanna (2016). "Maschinenlesbar — menschenlesbar. Über die grundlegende Ausrichtung der Edition". In: *Editio* 30.1, pp. 144–156.

Ramsay, Stephen (2011). *Reading Machines: Toward and Algorithmic Criticism*. University of Illinois Press.

Rasmussen, Carl E. and Christopher K. I. Williams (2006). *Gaussian Processes for Machine Learning*. Cambridge: MIT press.

Reese, Stephen D and Seth C Lewis (2009). "Framing the war on terror: The internalization of policy in the US press". In: *Journalism* 10.6, pp. 777–797.

Řehůřek, Radim and Petr Sojka (2010). "Software Framework for Topic Modelling with Large Corpora". In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50.

Reul, Christian, Christoph Wick, Uwe Springmann, and Frank Puppe (2017). "Transfer learning for OCRopus model training on early printed books". In: *Zeitschrift für Bibliothekskultur* 5, pp. 32–45.

Rosen-Zvi, Michal, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth (2004). "The author-topic model for authors and documents". In: *Proceedings of the Uncertainty in artificial intelligence conference*. 20, pp. 487–494.

Rudolph, Maja and David Blei (2018). "Dynamic embeddings for language evolution". In: *Proceedings of the Conference on World Wide Web*, pp. 1003–1011.

Rudolph, Maja, Francisco Ruiz, Stephan Mandt, and David Blei (2016). "Exponential family embeddings". In: *Advances in Neural Information Processing Systems*, pp. 478–486.

Ruiz, Javier (2011). *Access to the Agreement between Google Books and the British Library*. URL: `https://www.openrightsgroup.org/blog/access-to-the-agreement-between-google-books-and-the-british-library/`.

Rush, Alexander (2018). "The Annotated Transformer". In: *Proceedings of the Workshop for NLP Open Source Software at the Annual Meeting of the Association for Computational Linguistics*. 56, pp. 52–60.

Rusinek, Sinai and Nitzan Gado (2021). "Feeding a Gazetteer: Leveraging Word Embeddings for Toponym Mining". In: *Proceedings of the ACM SIGSPATIAL International Workshop on Geospatial Humanities*. 5, pp. 28–35.

Rust, Phillip, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott (2022). *Language Modelling with Pixels*. arXiv preprint, 2207.06991.

Rybicki, Jan and Magda Heydel (2013). "The stylistics and stylometry of collaborative translation: Woolf's Night and Day in Polish". In: *Literary and Linguistic Computing* 28.4, pp. 708–717.

Samek, Wojciech, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller (2021). "Explaining Deep Neural Networks and

Beyond: A Review of Methods and Applications". In: *Proceedings of the IEEE* 109.3, pp. 247–278.

Samek, Wojciech, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Vol. 11700. Springer Nature.

Schlitz, Stephanie (2014). "Digital Texts, Metadata, and the Multitude: New Directions in Participatory Editing". In: *Variants* 11, pp. 71–89.

Schmidt, Desmond (2016). "Using standoff properties for marking-up historical documents in the humanities". In: *Information Technology: Human Computation* 58, pp. 63–69.

Schöch, Christof (2017). "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama". In: *Digital Humanities Quarterly* 11.2, online.

Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke (2020). "Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen". In: *Zeitschrift für digitale Geisteswissenschaften*, online.

Sennrich R; Volk, Martin (2011). "Iterative, MT-based sentence alignment of parallel texts". In: *Proceedings of the Nordic Conference of Computational Linguistics*. 18, pp. 175–182.

Shakespeare, William (1833). *Shakespeare's dramatische Werke, übersetzt von August Wilhelm Schlegel, ergänzt und erläutert von Ludwig Tieck*. Verlag Georg Reimer.

Shillingsburg, Peter (2014). "Development Principles for Virtual Archives and Editions". In: *Variants* 11, pp. 9–28.

Shoemark, Philippa, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray (2019). "Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*. 9, pp. 66–76.

Siemens, Ray, Meagan Timney, Cara Leitch, Corina Koolen, and Alex Garnett (2012). "Toward modeling the social edition: An approach to understanding the electronic scholarly edition in the context of new and emerging social media". In: *Literary and Linguistic Computing* 27.4, pp. 445–461.

Sigg, C., B. Fischer, B. Ommer, V. Roth, and J. Buhmann (2007). "Non-Negative CCA for Audio-Visual Source Separation". In: *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing*, pp. 253–258.

Springmann, Uwe, Christian Reul, Stefanie Dipper, and Johannes Baiter (2018). "Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin". In: *Journal for Language Technology and Computational Linguistics. Special issue on automatic text and layout recognition* 33.1, pp. 97–114.

Strubell, Emma, Patrick Verga, David Belanger, and Andrew McCallum (2017). "Fast and Accurate Entity Recognition with Iterated Dilated Convolutions". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 2670–2680.

Sugiyama, Masashi, Matthias Krauledat, and Klaus-Robert Müller (2007). "Covariate shift adaptation by importance weighted cross validation". In: *Journal of Machine Learning Research* 8.5, pp. 985–1005.

Szymanski, Terrence (2017). "Temporal Word Analogies: Identifying Lexical Replacement with Diachronic Word Embeddings". In: *Proceedings of the Association for Computational Linguistics*. Vol. 2. 55, pp. 448–453.

Tahmasebi, Nina, Lars Borin, and Adam Jatowt (2018). *Survey of computational approaches to lexical semantic change*. arXiv preprint, 1811.06278.

Tsvetkov, Yulia, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer (2015). "Evaluation of word vector representations by subspace alignment". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 2049–2054.

Underwood, Ted (2014). "Understanding genre in a collection of a million volumes". In: *White Paper of the Digital Humanities Start-up Grant, University of Illinois, Urbana-Champaign*.

– (2022). "Mapping the Latent Spaces of Culture". In: *Startwords* 3, online.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30, online.

Vries, Terrance de, Ishan Misra, Changhan Wang, and Laurens van der Maaten (2019). "Does Object Recognition Work for Everyone?" In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Workshops*, pp. 52–59.

Wainwright, Martin J. and Michael I. Jordan (2008). "Graphical models, exponential families, and variational inference". In: *Foundations and Trends in Machine Learning* 1.1–2, pp. 1–305.

Wallach, Hanna M. (2006). "Topic Modeling: Beyond Bag-of-Words". In: *Proceedings of the International Conference on Machine Learning*. 23, pp. 977–984.

Witkowski, Georg (1924). *Textkritik und Editionstechnik neuerer Schriftwerke*. Leipzig: Verlag H. Haessel.

Xuan, Junyu, Jie Lu, Guangquan Zhang, and Xiangfeng Luo (2015). "Topic Model for Graph Mining". In: *Transactions on Cybernetics* 45.12, pp. 2792–2803.

Yao, Zijun, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong (2018). "Dynamic Word Embeddings for Evolving Semantic Discovery". In: *Proceedings of the ACM International Conference on Web Search and Data Mining*. 11, pp. 673–681.

Yu, Juntao, Bernd Bohnet, and Massimo Poesio (2020). "Named Entity Recognition as Dependency Parsing". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 58, pp. 6470–6476.

Zeng, Ziqian, Yichun Yin, Yangqiu Song, and Ming Zhang (2017). "Socialized Word Embeddings". In: *Proceedings of the International Joint Conference on Artificial Intelligence*. 26, pp. 3915–3921.

Zhang, Yating, Adam Jatowt, Sourav S Bhowmick, and Katsumi Tanaka (2016). "The past is not a foreign country: Detecting semantically similar terms across

time". In: *IEEE Transactions on Knowledge and Data Engineering* 28.10, pp. 2793–2807.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang (2017). "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 2979–2989.

Appendix A

Software Architecture Design of the Standoff Converter

In this section, the details about the main components of the Standoff Converter package are presented and the software design decisions are justified that were made during development.

The Standoff class The main component of the package is the Standoff class which connects the tree-representation of lxml with a standoff representation in table format. It has a `Standoff.tree` attribute where one can directly access the lxml tree and it has a `Standoff.table` property where one can access the standoff table. The standoff table has character granularity. As an example let's consider the following document:

```
1 <TEI><text>
2     <body>
3         <p>Let's start a novel. By writing a<lb/>sentence. </p>
4         <p>Or two</p>
5     </body>
6 </text></TEI>
```

Listing A.1: A simple example of a TEI document.

An excerpt of the standoff table is shown in Table A.1

pos	type	el	depth	text
0	open	text	0	-
0	text	-	-	\n
...	...	...	...	...
14	open	p	2	-
14	text	-	-	L
...	...	...	...	...
47	empty	lb	3	-

Table A.1: An excerpt from the standoff table of the following Code Listing A.1.

Each row is either of type 'open', 'close', 'empty' or 'text'. A text row does not have a depth or an element but it contains a single character. An element row has a link to the etree element ('el') and a depth value. The position column represents where the row is relative to the raw string of the document. The idea is that the two representations are isomorphic. When the API of the package is used to add or remove an element both data structures are changed and kept in sync. In the example of the previous section (Listing A.2), internally, first, the according rows are added to the table and afterwards within the lxml etree the direct parent element of the newly added sentence tag is identified and the subtree is rebuilt from the data of the standoff table.

```
so.add_inline(
    begin=start_ind,
    end=end_ind,
    tag="s",
    depth=None,
    attrib={'id':f'{isent}'}
)
```

Listing A.2: Function call for creating a new standoff annotation.

When applying modifications, the decision to maintain both data structures synchronized adds a certain amount of complexity and performance cost and there has been a considerable amount of effort spent into improving the performance of the Standoff Converter. A performance profiling with a variety of synthetic documents has been added at `tests/profiling`. It tests for cases with long and short documents, with shallow and deep documents, modifying the document deep down toward the leaves or close to the root. At the same time, having two different data structures available can be a huge performance gain for reading access. Also being able to choose between querying a tabular data structure or an xml tree with, for example, xpath increases accessibility of the Standoff Converter for a wider range of users.

The View class The View class takes as input a Standoff object and it creates a tabular data structure that keeps track of all modifications that are performed to create the plain text output. The modifications, such as `insert_tag_text` that inserts a specific character for an XML tag can be chained. In the following example, there are multiple modifications chained together, as shown in Listing A.3.

Listing A.3: Chaining of filters on the view object to yield the specific, desired plain text variant.

```
view = (
    View(so)
    .insert_tag_text("pb", ' ')
    .exclude_outside("body")
    .exclude_inside("note")
```

```
6       .exclude_inside("del")
7       .exclude_inside("abbr")
8       .remove_comments()
9       .shrink_whitespace()
10  )
```

This might be a common use case to focus on the text body with `exclude_outside ("body")` and also omit notes, deleted text and abbreviations, to remove comments and to shrink longer consecutive white spaces into a single one. The view object keeps two versions, the initial raw character string (immutable) and the modified one. This way, even conflicting operations can be applied. For example, first excluding a tag and then including it again will give the same output plain text as the initial view. When all modifications are performed, one can retrieve the plain text of the view by `view.get_plain()`. The view object should still be kept as it holds the reverse lookup to get back from character positions in the plain text version to table positions in the Standoff Converter. One should also not alter the view object after retrieving the plain text version as the reverse lookup of the view object matches this exact plain text version.

Appendix B

Enriching Humanities Data

B.1 Contracts between Google Books and Various Cultural Heritage Institutions

The contracts between

- a number of US-based libraries and Google is available here, https://web.archive.org/web/20120707144623/http:/thepublicindex.org/docs/libraries/cic.pdf
- the British Library and Google is available here, https://www.openrightsgroup.org/app/uploads/2020/03/BL-Google-Contract.pdf
- the National Library of the Netherlands and Google is available here, https://web.archive.org/web/20111025094345/http:/www.kb.nl/nieuws/2011/contract-google-kb.pdf
- the University of Michigan and Google is available here, http://web.archive.org/web/20050906002322/https:/www.lib.umich.edu/mdp/um-google-cooperative-agreement.pdf
- the University of Texas at Austin and Google is available here, https://web.archive.org/web/20151226021049/https:/www.lib.utexas.edu/sites/default/files/google/utexas_google_agreement.pdf
- the University of Virginia and Google is available here, https://web.archive.org/web/20120707144748/http:/thepublicindex.org/docs/libraries/virginia.pdf
- Scanning Solutions (for the Bibliotheque Municipale de Lyon) and Google is available here, https://web.archive.org/web/20120707144718/http:/thepublicindex.org/docs/libraries/lyon_ae.pdf
- University of California and Google is available here, https://web.archive.org/web/20120707144625/http:/thepublicindex.org/docs/libraries/california.pdf.

B.2 OCR Model Evaluation

In the third setting, multiple models were trained within each group, always training on all books of that group except one and using only the data of the left-out book for testing. In all settings, the performance of the off-the-shelf OCR model on the test set are reported for comparison.
As depicted in Table B.1, the performance of fine tuning improves character accuracy each time even for the held-out book. This shows that the fine-tuned model indeed did not overfit on a specific book but captures patterns of a specific script. It should be noted, that in some cases of the third experiment different volumes occur as individual samples, for example, the second volume of Anne of Geierstein by Scott was not held-out when tested for the third volume of Anne of Geierstein. Scripts in different volumes are often more similar than scripts of the same font type which might improve the outcome of this experiments in some cases.

Left-out identifier	base model	# train	# test	baseline acc.	fine-tuned acc.	δ
chroniclesofcano03scot	Antiq.	686	50	99.22	99.59	0.37
H9UwAQAAMAAJ	Frak,	3794	96	96.74	99.57	2.83
aNQwAQAAMAAJ	Frak,	3822	65	97.0	99.53	2.53
chroniclesofcano02scot	Antiq.	709	25	99.02	99.51	0.49
zDTMtgEACAAJ	Frak,	3794	96	95.05	99.43	4.38
anneofgeierstein03scot	Antiq.	708	26	98.68	99.34	0.66
t88yAQAAMAAJ	Frak,	3786	105	91.13	99.28	8.15
anneofgeierstein02scot	Antiq.	684	53	98.3	99.27	0.97
DNUwAQAAMAAJ	Frak,	3794	96	95.26	99.01	3.75
D5pMAAAAcAAJ	Frak,	3780	111	93.69	99.01	5.32
3pVMAAAAcAAJ	Frak,	3777	115	94.68	98.99	4.31
zviTtwEACAAJ	Frak,	3806	83	95.76	98.97	3.21
8AQoAAAAYAAJ	Frak,	3800	89	94.7	98.9	4.2
1VUJAAAAQAAJ	Antiq.	635	107	96.88	98.8	1.92
AdiKyqdlp4cC	Frak,	3793	97	92.34	98.47	6.13
rDUJAAAAQAAJ	Antiq.	639	103	97.85	98.42	0.57
quentindurward02scotuoft	Antiq.	687	49	97.35	98.34	0.99
HCRMAAAAcAAJ	Frak,	3739	157	91.28	98.28	7.0
J4knAAAAMAAJ	Antiq.	708	26	97.15	98.07	0.92
2jMfAAAAMAAJ	Frak,	3703	197	92.43	98.04	5.61
XtEyAQAAMAAJ	Frak,	3783	108	87.69	97.59	9.9
quentindurward01scotuoft	Antiq.	708	26	96.38	97.13	0.75
wggOAAAAQAAJ	Antiq.	710	24	92.52	96.89	4.37
_QgOAAAAQAAJ	Antiq.	664	75	94.43	96.66	2.23
fAoOAAAAQAAJ	Antiq.	685	51	94.72	96.61	1.89
4zQfAAAAMAAJ	Frak,	3701	199	88.68	96.37	7.69
PzMJAAAAQAAJ	Antiq.	662	77	90.7	95.49	4.79
u4cnAAAAMAAJ	Frak,	3795	95	91.31	95.21	3.9
7JVMAAAAcAAJ	Frak,	3780	112	71.35	94.62	23.27
8dAyAQAAMAAJ	Frak,	3780	111	84.45	94.24	9.79
htQwAQAAMAAJ	Frak,	3792	98	88.42	94.14	5.72
YAZXAAAAcAAJ	Frak,	1909	2190	80.68	92.92	12.24
MzQJAAAAQAAJ	Antiq.	691	45	84.9	89.52	4.62
kggOAAAAQAAJ	Antiq.	685	51	85.64	87.56	1.92
Fy4JAAAAQAAJ	Antiq.	709	25	78.9	85.15	6.25
oNEyAQAAMAAJ	Frak,	3798	92	66.31	84.79	18.48

Table B.1: Model performance evaluated with a leave-one-out strategy. Within each group (German Fraktur and English Antiqua), an individual model is trained on all samples except from the left-out identifier on which the model is tested afterwards. The performance of the fine-tuned model is improved in each case, often by a large margin (Lassner, Coburger, et al. 2022).

Appendix C

Word2Vec with Structure

C.1 Implementation Details

C.1.1 Ex1

All word embeddings were trained with $d = 50$.

GloVe We run GloVe experiments with $\alpha = 100$ and minimum occurrence = 25.

Skip-Gram, CBOW We use the Gensim Řehůřek and Sojka 2010 implementation of Skip-Gram and CBOW with min_alpha = 0.0001, sample = 0.001 to reduce frequent words and for Skip-Gram, we use 5 negative words and ns_component = 0.75.

Parameter selection The parameters λ and τ for DW2V, W2VConstr and W2VPred were selected based on the performance in the analogy tests on the train set. In order to flatten the contributions from the n nearest neighbors (for $n = 1, 5, 10$), we rescaled the accuracies: For each n, accuracies are scaled so that the best and the worst method is 1 and 0, respectively. Then, we computed their average and maximum.

Analogies Each analogy consists of two word pairs (e.g., countryA - capitalA; countryB - capitalB). We estimate the vector for the last word by $\widehat{v}$ = capitalA - countryA + countryB, and check if capitalB is contained in the n nearest neighbors of the resulting vector $\widehat{v}$.

C.1.2 Ex2

Temporal Analogies Each of two word pairs consists of a year and a corresponding term, as e.g., 2000 - Bush; 2008 - Obama, and the inference accuracy of the last word by vector operations on the former three tokens in the embedded space is evaluated. To apply these analogies, GloVe, Skip-Gram and CBOW are trained individually on each year on the same vocabulary as W2VPred (same parameters for

GloVe as before, with minimum occurrence=10). For the other methods, DW2V, W2VConstr, and W2VPred, we can simply use the embedding obtained in Section 5.3.3. Note that the parameters τ and λ were optimized based on the general analogy tests.

C.1.3 Ex3

Burrows It compares normalized bag-of-words features of documents and subcorpora, and provides a distance measure between them. Its parameters specify which word frequencies are taken into account. We found that considering the 100th to the 300th most frequent words gives the best structure prediction performance on the train set.

Recall@k Let $\hat{D} \in \mathbb{R}^{T \times T}$ be the predicted structure. We report on recall@k averaged over all domains:

$$\text{recall@k} = \tfrac{1}{T} \sum_t^T \text{recall@k}_t, \qquad \text{where}$$

$\text{recall@k}_t = \frac{\text{TP}_t(k)}{\text{TP}_t(k)+\text{FN}_t(k)}$,
$\text{TP}_t(k) = \sum_{t'}^T b(D_t, t', k)$ & $b(\hat{D}_t, t', k)$,
$\text{FN}_t(k) = \sum_{t'}^T b(D_t, t', k)$ & $\neg b(\hat{D}_t, t', k)$, and

$$b(x, i, k) = \begin{cases} 1 & x_i \text{ is one of the k smallest in } x, \\ 0 & \text{otherwise.} \end{cases}$$

For NYT, we chose $k = 2$, which means relevant nodes are the two next neighbors, i.e., the preceding and the following years. For WikiFoS and WikiPhil, we respectively chose $k = 3$ and $k = 2$, which corresponds to the number of subcategories that each main category consists of.

W2VPred Hyperparameters for W2VPred were selected on the train set where we maximized the accuracy on the global analogy test as before.

C.2 Preprocessing of the Datasets

We lemmatized all tokens, i.e., assigned their base forms with spacy[1] and grouped the data by years (for NYT) or categories (for WikiPhil and WikiFoS). For each dataset, we defined one individual vocabulary where we considered the 20,000 most frequent (lemmatized) words of the entire dataset that are also within the 20,000 most frequent words in at least 3 independent slices, i.e., years or categories. This way, we filtered out "trend" words that are of significance only within a very short time period/only a few categories. The 100 most frequent words were filtered out

[1] See `https://spacy.io`.

Dataset	ρ
NYT	0.58
WikiFoS	0.65
WikiPhil	-0.19
WikiPhil(denoised)	-0.14

Table C.1: Pearson correlation coefficients for performance on analogy tests ($n = 10$) and structure prediction evaluation (recall@k) by W2VPred for the parameters applied in the grid search. Linear correlation indicates that a good word embedding quality also leads to an accurate structure prediction (and vice versa). Significant correlation coefficients ($p < 0.05$) are marked in gray.

as stop words. We set the symmetric context window (the number of words before and after a specific word considered as context for the PPMI matrix) to 5.

C.3 Assessment of Prior Structure

In the following, we reevaluate the aforementioned prior affinity matrix for WikiPhil which is depicted in Figure 5.1. Therefore, we analyse the correlation between embedding quality and structure performance and find that a suitable ground truth affinity matrix is necessary to train good word embeddings with W2VConstr. We trained W2VPred with different parameter settings for (λ, τ) on the train set, and applied the global analogy tests and the structure prediction performance evaluation (with the prior structure as the ground truth). For λ and τ, we considered log-scaled parameters in the ranges $[2^{-2}-2^{12}]$ and $[2^{4}-2^{12}]$, respectively, and display correlation values on NYT, WikiFoS, and WikiPhil in Table C.1.
In NYT and WikiFoS, we observe clear positive correlations between the embedding quality and the structure prediction performance, which implies that the estimated structure closer to the ground truth enhances the embedding quality. The Pearson correlation coefficients are 0.58 and 0.65, respectively (both with $p < 0.05$).
Whereas Table C.1 for WikiPhil does not show a clear positive correlation. Indeed, the Pearson correlation coefficient is even negative with -0.19 which implies that the prior structure for WikiPhil is not suitable and even harmful for the word embedding performance. This result is consistent with the bad performance of W2VConstr on WikiPhil in Sub-Section 5.3.3. After evaluating the newly discovered structure, the correlation between the embedding quality and the structure prediction performance–with the denoised estimated affinity matrix as the ground truth–is shown in Table C.1. The Pearson correlation is still negative -0.14 but not statistically significant anymore ($p = 0.11$).

Appendix D

AlterLDA

D.1 Derivation of the Collapsed Gibbs Sampler

For the Collapsed Gibbs Sampler of the alterLDA model it is shown how to derive the posterior for the topic assignment at a current position, given the current configuration. First, the joint probability of the whole model is given before showing how to compute the topic assignment based on count statistics. The joint probability of the model ist given by

$$\begin{aligned} p(\mathbf{w}, c, z, \gamma, \beta, \theta \mid \alpha, \eta, \xi) =& p(c \mid z) \cdot p(\mathbf{w} \mid z, \beta) \cdot p(\gamma \mid \xi) \cdot p(\beta \mid \eta) \cdot p(z \mid \theta) \cdot p(\theta \mid \alpha) \\ =& \prod^{M,N} \text{cat}(c \mid c, \gamma) \times \prod^{M,N} \text{cat}(\mathbf{w} \mid z, \beta) \times \prod^{M,N} \text{cat}(z \mid \theta) \\ & \times \prod^{M} \text{dir}(\theta \mid \alpha) \times \prod^{K} \text{dir}(\gamma \mid \xi) \times \prod^{K} \text{dir}(\beta \mid \eta) \end{aligned}$$

We introduce a counter variable c which can be indexed in four dimensions, the current topic (k), the current document (m), the current alteration mode (a) and the current token ($\mathbf{w}$).

$$c_{k,m,a,\mathbf{w}} = \sum_{n=1}^{N} \mathbf{I}(z_{m,n} = k \quad \& \quad w_{m,n} = \mathbf{w} \quad \& \quad c_{m,n} = a)$$

In this setting, the desired computation is the probability of a topic assignment at a specific position given a current configuration of all other topic assignments. This probability can be formalized by

$$p(z_{m,n} \mid z_{-(m,n)}, \mathbf{w}, c, \alpha, \eta, \xi) \propto \; p(z_{m,n}, z_{-(m,n)}, \mathbf{w}, c \mid \alpha, \eta, \xi)$$

Adopting Equation 16 from Carpenter 2010, this probability can be written by marginalizing θ, β and γ from the joint probability.

Identifier	Explanation	Type	Size
V	Number of unique tokens in the dictionary	Int	
W	Number of tokens in the corpus	Int	
M	Number of documents	Int	
Nm	Number of tokens in document m	Int	
K	Number of topics	Int	
α	Concentration of θ	Hyper parameter	K
η	Concentration of β	Hyper parameter	V
ξ	Concentration of γ	Hyper parameter	2
β	Topic-term variable	Dirichlet	K x V
θ	Document-topic variable	Dirichlet	M x K
γ	Topic-alteration-tendency variable	Dirichlet	K x 2
z	Token-topic variable	Categorical	W x K
w	Tokens	Observed (Categorical)	W x V
c	Alteration	Observed (Categorical)	W X 2

Table D.1: Documentation of the symbols used for the AlterLDA method.

$$
\begin{aligned}
p(z_{m,n}, z_{-(m,n)}, \mathbf{w}, c \mid \alpha, \eta, \xi) &= \int\int\int p(\mathbf{w}, c, z, \gamma, \beta, \theta \mid \alpha, \eta, \xi) d\theta \; d\beta \; d\gamma \\
&= \underbrace{\int p(\theta \mid \alpha) \cdot \; p(z \mid \theta) d\theta}_{A} \\
&\quad \times \underbrace{\int p(\mathbf{w} \mid z, \beta) \cdot \; p(\beta \mid \eta) d\beta}_{B} \\
&\quad \times \int p(c \mid z, \gamma) \cdot \; p(\gamma \mid \xi) d\gamma \\
&= \prod_{k=1}^{K} \int p(\gamma_k \mid \xi) \prod_{m=1,n=1}^{M,N_m} p(c \mid \gamma_{\text{argmax}(z_{m,n})}) d\gamma_k \times A \times B
\end{aligned}
$$

A and B are substitued here because their derivation is identical to the one in Carpenter 2010 Analogue to Equation 27 of Carpenter 2010, after inserting the definitions of the Dirichlet distribution the result is proportional to three factors.

$$
\propto (\mathbf{c}^{-}_{z_{m,n},*,*,*} + \alpha_{z_{m,n}}) \left(\frac{\mathbf{c}^{-}_{z_{m,n},*,*,\mathbf{w}_{m,n}} + \eta_{w_{m,n}}}{\mathbf{c}^{-}_{z_{m,n},*,*,*} + \sum_v^V \eta_v} \right) \left(\frac{\mathbf{c}^{-}_{z_{m,n},*,c_{m,n},*} + \xi_{w_{m,n}}}{\mathbf{c}^{-}_{z_{m,n},*,*,*} + \sum_i^2 \xi_i} \right)
$$

Where $\cdot^{-}$ denotes the counter disregarding the current position n, m.

D.2 Non-content-related Alteration Processing

To identify the non-content-related alteration categories, existing methods are used, however, for the identification of the specific content related reasons on the right, the novel Alteration Latent Dirichlet Allocation (AlterLDA) method is used. As the main contribution of this work lays in introducing the new alterLDA model, the description of the non-content-related alterations. In this section, the description of the established methods is shortly summarized, whereas the description of alterLDA is given more space in the forthcoming subsections.

D.2.1 Paratexts

In Figure 6.8, the excerpt of the facsimile marked as archival note (orange) has the number 6 written in the top right corner of the sheet, this detail is shown in Figure D.1. The corresponding xml transcription is the following:

```
<note type="foliation" place="margin-right inline"
                        hand="#pencil_1">6</note>
```

As the header reveals, this pencil numbering has been performed by an archivist:

```
<handNote xml:id="pencil_1" scope="minor" medium="pencil"
                        scribe="archivist">
       <seg xml:lang="de">Hand eines Archivars, in
                        Bleistift.</seg>
       <seg xml:lang="en">Hand of an archivist, in
                        pencil.</seg>
       <seg xml:lang="fr">Main d'un archiviste, crayon de
                        papier.</seg>
</handNote>
```

For the second pencil note in Figure D.1 there is no scribe annotated although it also contains nothing but a numbering:

```
<note type="foliation" place="align(center)" hand="#pencil_2">
       <hi rend="underline">99</hi>
</note>
```

The additions that have been performed by a different hand than the primary author and that contain numberings or dates, we consider to be archivists notes. Such archivist's or editor's additions can be identified with a very basic set of rules.

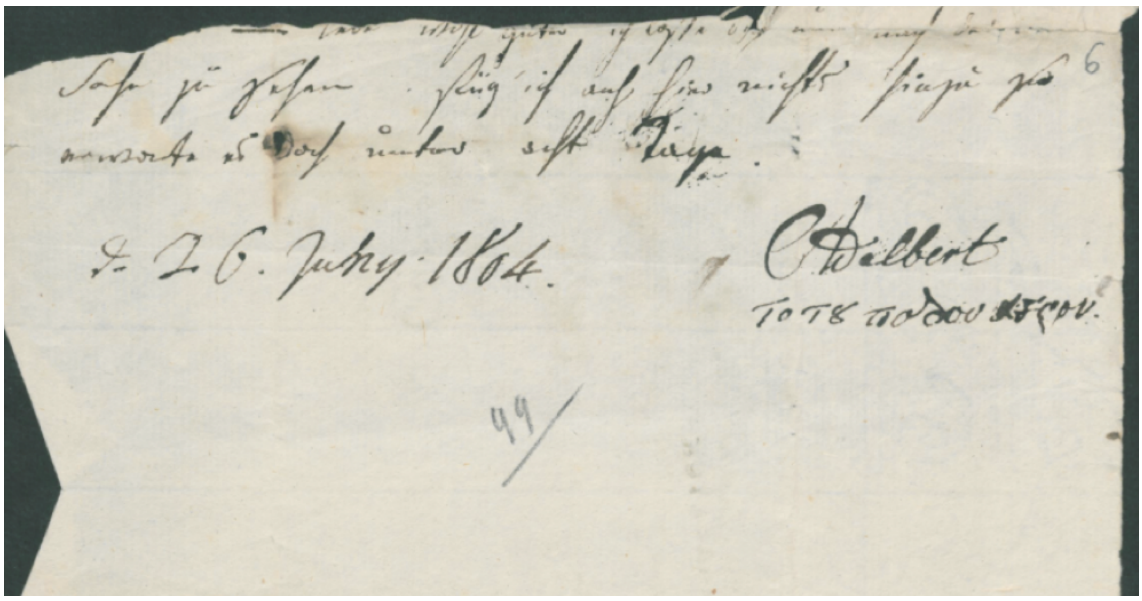

Figure D.1: Archival note. BI, Adelbert von Chamisso to Louis de la Foye. Nachlass 239, Blaat 6. Staatsbibliothek Berlin / Manuscripts section. Reuse subject to prior approval by Staatsbibliothek Berlin. Published in; Letter from Adelbert von Chamisso to Louis de La Foye (fragment) (without place, 26 june 1804). Ed. by Anna Busch, Sabine Seifert. Prepared by Janine Katins. In collaboration with Sabine Seifert, Sophia Zeil. In: "Letters and texts: Intellectual Berlin around 1800." Ed. by Anne Baillot. Berlin: Humboldt-Universität zu Berlin. `http://www.berliner-intellektuelle.eu/manuscript?Brief005ChamissoandeLaFoye`. Last modified: 27 April 2015.

D.2.2 Corrections

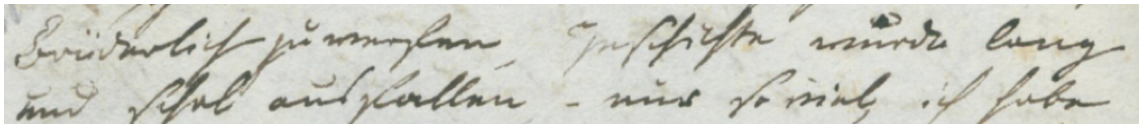

Figure D.2: Correction of mistake of "wurde" to "würde". BI, Adelbert von Chamisso to Louis de La Foye. Nachlass 239, Blatt 85. Staatsbibliothek Berlin / Manuscripts section. Reuse subject to prior approval by Staatsbibliothek Berlin. Published in: Letter from Adelbert von Chamisso to Louis de La Foye (Geneva, at the beginning of 1812). Ed. by Anna Busch, Sabine Seifert. Prepared by Lena Ebert. In: "Letters and texts: Intellectual Berlin around 1800". Ed. by Anne Baillot. Berlin: Humboldt-Universität zu Berlin. http://www.berliner-intellektuelle.eu-/manuscript?Brief047ChamissoandeLaFoye. Last modified: 27 April 2015.

The alteration marked in green replaces a single character of a word, for which the corresponding part of the facsimile is shown in Figure D.2. "[..]Geschichte wuürde lang und schal ausfallen[..]" BI, Adelbert von Chamisso to Louis de La Foye. Letter

47, p. 1 This alteration is a correction of a mistake and conceptually, it is worth noting that the words before and after the alteration are very similar. This characteristic will be exploited for the identification of corrections. Identifying corrections is a considerably more difficult task because the corrected version does not necessarily have to be correct from what we know today. The fact that the alteration author corrected the text only means that he or she thought that his or her version is correct. We thus cannot rely on comparing the second version of the text with what an automatic spell checker would the first version correct to. Instead, we divided the problem even further into spelling alterations and grammatical alterations. For identifying spelling mistakes, the tokens of both versions are fuzzy-string matched against the common dictionary of lemmas. If both tokens match closely to the same lemma according to the Levenshtein distance, the two tokens are considered two different spellings of the same word. Fuzzy string matching of multiple tokens against a large vocabulary can be costly in terms of computing time and memory. For a larger data set an adjustment to this approach may be necessary. However, this approach gave better results than simply comparing Levenshtein distance of the tokens of both versions with each other, due to smaller tokens that are very similar but mean different things (e.g. “hate” and “fate” have Levenshtein distance of 1 but have a very different meaning.) For identifying grammatical alterations, we assume that the forms of the tokens in the sentence change and probably punctuations are added or deleted, but the set of lemmas is preserved for the most part. Hence, if the forms or the part of speech of the tokens in the span change but the set of lemmas do not, this alteration is a grammatical correction.

D.2.3 Stylistic Alterations

For identifying stylistic alterations, we assume that all corrections and paratexts are already labelled according to the described method. Thus, there are only stylistic alterations and moral censorships left to be labelled. In our understanding, a stylistic alteration preserves the meaning of the text by only changing the way it is posed which includes rearranging of words, the use of synonyms and rephrasing. In recent years, a method gained a lot of attention that strives to find a vector-space representation of words that capture its meaning. Words or sentences projected to this space reveal a high similarity (for example cosine-similarity) if they have the same meaning. We introduce a threshold and consider all alterations for which the vector-space embedding of the text before and after the alteration reveal a smaller distance to be a stylistic alteration.
The alteration marked in blue (Figure 6.8) which is shown in higher resolution in Figure D.3, reorders the words at the beginning of a sentence without changing the meaning.

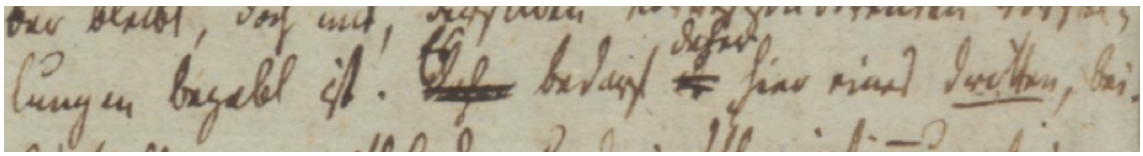

Figure D.3: Stylistic alteration of “Daher bedarf es” to “Es bedarf daher”. BI, About the notion of philosophy PP. by Immanuel Hermann von Fichte, p. 14.

For completeness, we also provide the individual facsimile of the content related alteration example in Figure D.4.

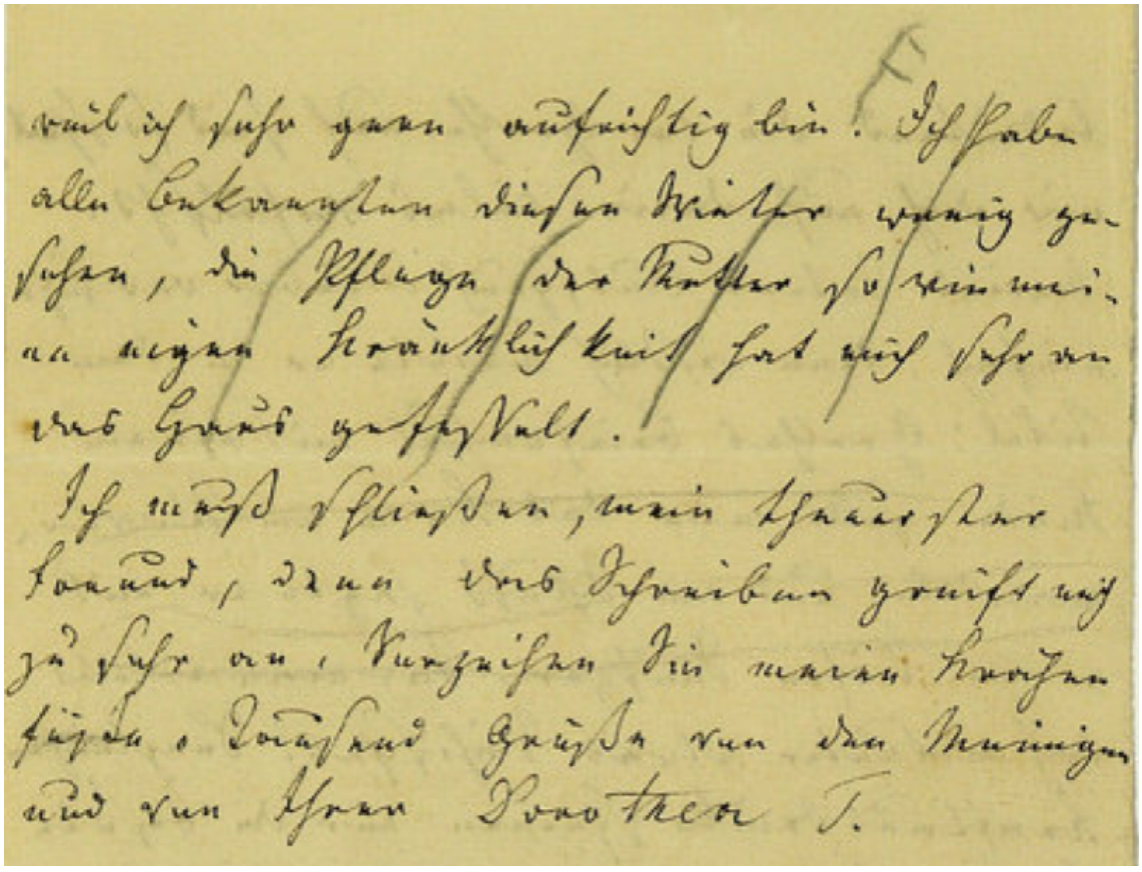

Figure D.4: Content-related alteration. Nachlass Uechtritz. Oberlausitzische Bibliothek der Wissenschaften Görlitz. Reuse subject to prior approval by Oberlausitzische Bibliothek der Wissenschaften Görlitz. Letter from Dorothea Tieck to Friedrich von Uechtritz (Dresden, 10 April 1835). Ed. by Sophia Zeil. Published in: “Letters and texts: Intellectual Berlin around 1800.” Ed. by Anne Baillot. Berlin: Humboldt-Universität zu Berlin. `http://www.berliner-intellektuelle.eu/manuscript?Brief16DorotheaTieckanUechtritz`. Last modified: 24 January 2015.

D.3 Results on Synthetic Data

A big advantage of generative models is that they can be used to generate new data. The generated documents themselves may not be too interesting in the case of topic models, but they can be used to evaluate the functionality of the model. To do this, first, the variables of the model are initialized, and documents are generated. Now the variables are initialized again and based on the previously generated documents the old variable configurations are reconstructed. The performance of the inference can be measured by the accuracy of the reconstruction.
In Figure D.5, the results of such an evaluation over 324 different experiment runs is shown, within the sparsity of the hyper-parameters as well as the number of tokens were varied. Each cell shows the mean reconstruction of c over two runs with a given set of parameter choices. The brighter the color of the cell, the better the reconstruction. Overall, a smaller alpha yields better results, independent of the size of the data set and the choice of the other concentration factors. Interestingly, for fewer documents and $\alpha = 0.1$ a smaller concentration factor η performs better, whereas for either a larger number of documents or a larger $\alpha = 1.0$, a larger η is to be preferred. The explanation for this result is that with more documents, the exact proportions of the topics can be inferred more accurately, whereas for fewer documents, there is the chance of getting a few (sparse) topics right.

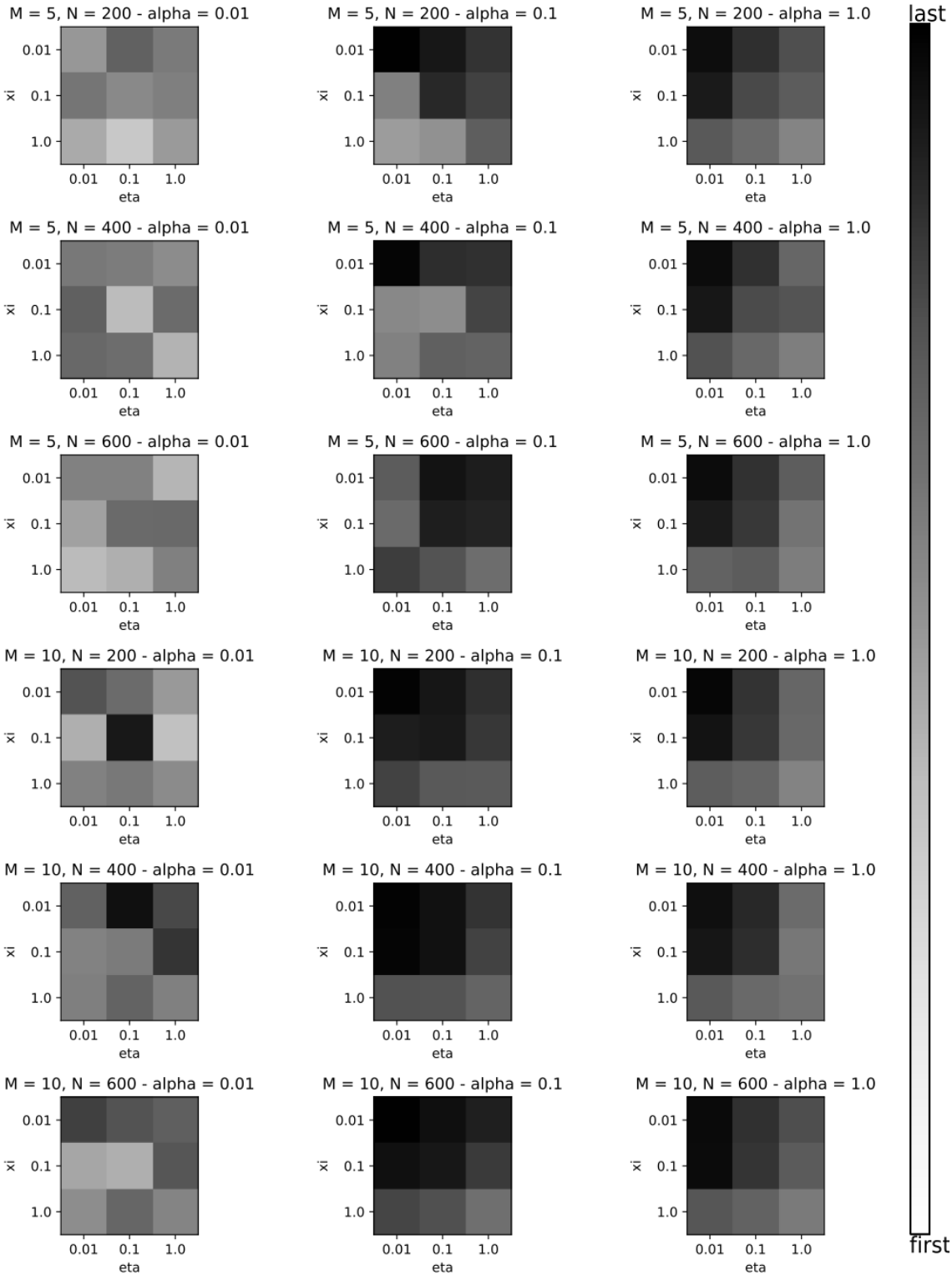

Figure D.5: Grid search result for training accuracy of the $\hat{c}$ parameter on synthetic data. Even with a small total number of tokens, the accuracy can be very high. Interestingly, the accuracy depends strongly on the sparsity of α, ξ, and η.