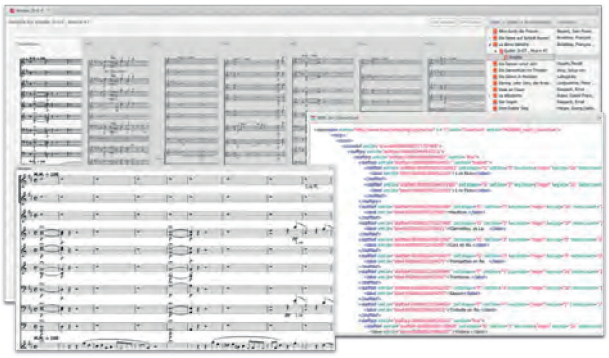

Abbildung 3: Darstellung der Incipits

Eine weitere Besonderheit des Projekts ist die exemplarische sog. Tiefenerschließung einiger ausgewählter Aufführungsmaterialien: Bei diesen werden auch die Faksimiles der Quellen zur Verfügung gestellt und zwar in einer Aufbereitung für einen taktgenauen Zugriff. Nur durch diese Form der Erschließung ist es z. B. möglich, Eingriffe in den Notentext nicht auf Grund der Materialität (also z. B.: Streichung auf Bl 4v bis 5r vorletzter Takt), sondern inhaltlich (z. B.: Streichung in Nr. 1 von T. 17–20) und damit für den Nutzer (mit Hilfe anderer Materialien, also gedruckter oder anderer handschriftlicher Quellen) nachvollziehbar anzugeben. Zur Erstellung der sog. Vertaktung wird die Software „Edirom" benutzt und zur Darstellung ist das „Theatre Tool" mit Edirom Online (https://github.com/Edirom/Edirom-Online) verknüpft. Diese Software wurde zwar für die Aufbereitung von Notenmaterial entwickelt, aber es lassen sich damit auch Textquellen z. B. nach Szenen oder sogar Zeilen kartieren.

Das Theatre-Tool ist für die Darstellung dieser komplexen Text- und Datenstrukturen entwickelt, kann aber leicht an andere Anforderungen angepasst werden: Bei dem im Projekt erfassten Material handelt es sich z. B. überwiegend um handschriftliches Material, weshalb die nach FRBR vorgesehene vierte Ebene, das Exemplar (item), nach der Regel der „manifestation singleton" nicht berücksichtigt wird. Selbstverständlich wäre aber auch diese darstellbar. Da das Hauptinteresse der Erschließung auf der Arbeitsweise und dem Personal der Detmolder Hoftheater-Gesellschaft liegt, werden die erwähnten Orte zwar ausgezeichnet, gibt es für diese aber keine eigenständigen Dateien (mit der Möglichkeit zu Referenzen) und bislang keine Suchmöglichkeit.

Mit der zunehmenden Digitalisierung der Bestände durch die Bibliotheken könnten diese über iiiF in das Theatre Tool eingebunden werden, wodurch etliche rechtliche Probleme gelöst werden könnten. Wie damit auch eine Vertaktung verbunden werden kann, wäre zu überprüfen.

Weiterer Abstimmungsbedarf, an dem aber beidseitig großes Interesse besteht, ist notwendig zwischen Wissenschaft und Bibliothek. Es ist selbstverständlich, dass die Beispiele der Tiefenerschließung des Projekts ebenso wie die Erstellung z. B. von Komponisten-Werkverzeichnissen nur durch die Wissenschaft zu leisten sind. Dennoch besteht großes Interesse, diese Detailinformationen zu einzelnen Quellen auch über die besitzende Bibliotheken zugänglich zu machen Die Verwendung von Standards und Normdaten wie sie im Hoftheater-Projekt erprobt worden sind, bildet hierzu einen erster Schritt, doch muss sicherlich auch verstärkt über Schnittstellen für den Datenaustausch nachgedacht werden.

Bibliographie

Kamzelak, Roland S. (2016): „Digitale Editionen im semantic web. Chancen und Grenzen von Normdaten, FRBR und RDF" in: Richts, Kristina / Stadler, Peter (eds.): *„ei, dem alten Herrn zoll ich Achtung gern".* Festschrift für Joachim Veit zum 60. Geburtstag. München: Allitera Verlag 423–435; online unter: https://nbn-resolving.org/urn:nbn:de:bsz:14-qucosa2-233392

Münzmay, Andreas (2018): „Lesen und Schreiben im digitalen Dickicht, Musikwissenschaft, Digital Humanitites und die hybride Musikbibliothek" in: BIBLIOTHEK Forschung und Praxis 42; 236–246.

Münzmay, Andreas (2019): „Kulturtransferforschung und Musikwissenschaft", in: Calella, Michele / Leßmann, Benedikt (eds.): *Zwischen Transfer und Transformation: Horizonte der Rezeption von Musik* (= Wiener Veröffentlichungen zur Musikwissenschaft 51). Wien 175–190.

Richts, Kristina / Veit, Joachim (2018): „Stand und Perspektiven der Nutzung von MEI in der Musikwissenschaft und in Bibliotheken" in: BIBLIOTHEK Forschung und Praxis 42: 292–301.

Pernerstorfer, Matthias J. (2012): *Theater – Zettel – Sammlungen. Erschliessung, Digitalisierung, Forschung*. Wien (= Don Juan Archiv Wien: Bibliographica, 1)

Pernerstorfer, Matthias J. (2015): *Theater – Zettel – Sammlungen* Bd. 2: *Bestände, Erschließung, Forschung.* Wien 2015 (= Don Juan Archiv Wien: Bibliographica 2)

Veit, Joachim (2020): „Notistenspezifische Erwartungen der Wissenschaft an die Web-Präsentation digitalisierter Musikhandschriftenbestände" in *: Das Instrumentalrepertoire der Dresdner Hofkapelle in den ersten beiden Dritteln des 18. Jahrhunderts – Überlieferung und Notisten.*

Wiermann, Barbara (2018a): „Bibliothekarische Normdaten und digitale Musikwissenschaft" in: *Die Musikforschung*, 71: 338–357.

Wiermann, Barbara (2018b): „musiconn. performance: musikalische Ereignisdaten im Fachinformationsdienst Musikwissenschaft" in: Bonte, Achim / Rehnolt, Juliane (eds.): *Kooperative Informationsinfrastrukturen als Chance und Herausforderung*. Thomas Bürger zum 65. Geburtstag herausgegeben von. Berlin, Boston 398–415.

The rapid rise of Fraktur

Weichselbaumer, Nikolaus
nikolaus@weichselbaumer.info
JGU Mainz, Deutschland

Seuret, Mathias
mathias.seuret@fau.de
FAU Erlangen, Deutschland

Limbach, Saskia
limbach@uni-mainz.de
JGU Mainz, Deutschland

Hinrichsen, Lena
lhinrich@students.uni-mainz.de
JGU Mainz, Deutschland

Maier, Andreas
andreas.maier@fau.de
FAU Erlangen, Deutschland

Christlein, Vincent
vincent.christlein@fau.de
FAU Erlangen, Deutschland

Introduction

From the first experiments in 1513, Fraktur quickly became the most successful gothic font in print history. Whereas gothic fonts in most other countries went out of use in the 16th and 17th centuries, Fraktur became by far the most used font for German texts in the early modern period. The font also made it to modernity and was used frequently, almost unchanged, until the middle of the 20th century. Even today the font is often used especially when a design should appear 'historical'.

Despite its importance, fairly little is known about the famous font. The origins of Fraktur at the beginning of the 16th century and the possible creators Vincenz Rockner and Johann Neudörffer have been the subjects of several studies (Kautzsch 1922, Kapr 1993: 24, Hessel 1937). Apart from this, however, we know remarkably little about its development over the following centuries. Only the Antiqua-Fraktur dispute around 1800 gained the interest of book historians again when German intellectuals discussed which of the two fonts is more appropriate for German texts (Lühmann 1981, Killius 1999). Yet the emergence of Fraktur and its leading role in font history remains understudied.

Tracing the emergence of Fraktur is complicated by two facts: On the one hand, contemporary evidence, such as invoices, letters and type specimens, is at best fragmentary and nearly impossible to contextualise without an analysis of the books themselves. On the other hand, researchers are simply overwhelmed by the amount of material available. For the 16th century alone, the German national bibliography VD16 (www.vd16.de) lists over 100,000 titles. This makes it impractical to look at every book individually and determine its fonts or even only its main text font.

Recent research presents a solution to this problem. With the help of a newly developed pattern recognition tool, large amounts of digitised book pages can be categorised into font groups. This tool was developed in the context of a project on font-specific OCR (Weichselbaumer et al. 2019, Seuret et al. 2019) and was then used for a large dataset of digitised books from BSB Munich. This paper will present the results and provide new insights into the rapid rise of Fraktur.

Methodology

Our methodology is based on automatic document image labeling which is done by a deep convolutional neural network (CNN) trained for font classification. As artificial intelligence typically requires a great amount of data, we manually prepared a training dataset of more than 35'000 document images, each labeled with the used fonts. We recently published this dataset along with a complete description of the approach we used (Seuret et al. 2019). For these test pages, we reach an accuracy slightly higher than 98% for recognising the main font.

As CNN architecture, we employ a DenseNet-121 (Huang et al. 2017). It is composed of 121 neural layers, most of them contained in 4 densely connected blocks. To identify the main font in a document image, we split it into many overlapping 224x224 px large patches, which are subsequently passed to the CNN. The overall page result is obtained by taking the average of all classified patches. Processing pages patch-wise is significantly more memory-friendly than using fully-convolutional neural networks and does not require expensive hardware.

For this study, book processing was done in two steps. First, we extracted the production years and the language of digitised books from the available metadata. We disregarded books that were not tagged as German as well as those without a clear date of publication (using 15 processing rules for the dates, in addition to an extra-permissive roman numbers parser). Second, we identified the main font of the pages 10 to 19 of every book, thereby avoiding prefaces and title pages which can differ quite decisively from the rest of the book. In case the network did not detect the same font on at least 6 pages of the same book, we disregarded the entire 10 pages. This way we automatically labelled 10 pages of a total of 85'165 books as the basis for this study.

Library catalogues are a great resource for metadata. Yet, in many cases early printed books were collated differently, even within the same library. This is largely the result of changing bibliographical practices in the past decades in which only slowly a standard emerged. Therefore, we often find metadata that is not standardised.The date of publication is often far from straight-forward. It may just be an estimation with words like 'ca.', 'um' or 'etwa'; it may include two years, such as 1549/50; or it may be displayed in Roman numerals, which sometimes differ from modern-day practice, such as 'MDXXXX'.

In parsing this data, we attempted to keep as much usable data as possible without distorting the results. Roman numerals were transformed into arabic numbers. In case two years are given (e.g. 1549/50), the first and second year were alternated. If the year was given as a time span (e.g. 1650-1660) we computed the average of both values and rounded to the larger number when necessary. For the estimated dates we decided that omitting the records altogether would have shrunk the database considerably. So we just deleted the estimation markers like 'ca.' and kept the years. This produced larger spikes every 50 years and smaller ones every 5 and 10 years, but these can be explained easily when interpreting the results.

After we received the results of the network we double-checked unlikely results by hand. This included some 60 books printed after 1550 which were classified as fonts predominantly used in the incunabula era - Rotunda, Textura, Gotico-Antiqua and Bastarda. In most cases these were actually empty pages with text bleeding through the other side of the page. Occasionally there were also tables with arabic numerals which were classified wrongly as one of the fonts mentioned above. We decided to delete this small number of misclassified books.

Results

The resulting data[1], which shows the publication of books in the German language, seems to be fairly representative of the general print production in Germany. After a relatively slow start up to 1520, the Reformation led to a very considerable spike in print production. The Thirty Years War (1618-1648) brought the print industry almost to a standstill. After that we see a steady rise in the number of editions per year, quickly accelerating at the end of the 18th century. Interestingly, the slight drop towards the very end of the century is rather unexpected. It may just be the result of the library's preference to digitise material pre-1800.

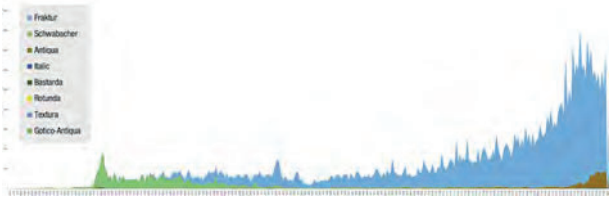

Figure 1: Main font groups in 85,223 digitised books from the Bavarian State Library, printed between 1472 and 1800[2]

With regard to font groups, the diagram showing absolute values (Fig. 1) stresses the fact that for German texts, Schwabacher and Fraktur are by far the two most important font groups. Yet, due to the much higher print production in the 18th century, Fraktur appears approximately 8.5 times as many times as a main font as Schwabacher.

In the results, you can also find a negligible number of Hebrew (4) and Greek (2) recognized as main fonts. They either actually aren't German (bsb10239978) pointing to a rare mistake in the metadata of the books, or contain pages with mainly Hebrew/Greek characters (bsb10779648 / bsb10360987). The book bsb11254779 (apparently written for Christians in Israel) is mainly Hebrew but has a German title. The *Catechismus D. Martini Lutheri minor: E lingua vernacula in Latinam & Graecam pridem translates* (bsb11229498) has been recognized as Greek although this is only true for 1/3 of the text.

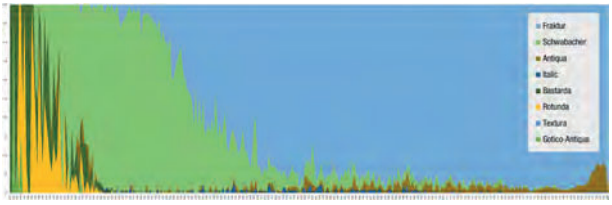

Figure 2: Main font groups in 85,223 digitised books from the Bavarian State Library, printed between 1472 and 1800, normalised in percentage

But absolute numbers only tell half the story. In order to know how important a font was at a given time, it is more fruitful to look at its share of the print production in a given year. In a normalised diagram of the same data we see much noise for the first decades of print up to the 1520ies. This is caused by the very low print production at that time and by the fact that printers often used fonts inconsistently as they did not have a complete set of font styles available. Nevertheless, the diagram shows that in the first decades of print, the two most important fonts for German texts seem to have been Bastarda and Rotunda. They were then gradually replaced by Schwabacher from the 1490ies onwards. Schwabacher reaches its largest share in the 1520ies to the 1540ies, the height of Reformation printing, before it is gradually replaced by a relatively new font - Fraktur. It is firmly established as the main font for German from about 1585 onwards. Only in the last decades of the 18th century does another font, Antiqua, become slightly more important in the production of German books. This indicates that the Antiqua-Fraktur debate had indeed some impact on contemporary book design. However, the overwhelming majority of books were still printed in Fraktur.

Conclusion

This study shows that Schwabacher dominated German language printing for the larger part of the 16th century until a fairly slow and linear change brought Fraktur to the dominant role it then kept through the 17th and 18th century. This makes the rise of Fraktur no less decisive, but significantly slower than often assumed (Kapr 1991, p 42; Killius 1999, p. 82).

These results have implications not only for the history of typography but also for OCR. When institutions use OCR engines, it is vital to choose the correct model for the specific text font. Quite commonly libraries use either Antiqua or Fraktur when a Schwabacher model or a mixed model could actually produce much better results, especially for books printed in the 16th century.

The used method promises to be a helpful and viable tool for digital book history. It paves the way for further studies on the statistical analysis of font use in early printed books and at the same time allows further research on the reasons for the change from Schwabacher to Fraktur. In addition, it offers the opportunity to shed more light on the role of type foundries in the development of book design in the early modern period.

Fußnoten

1. All data produced for this paper can be downloaded here: https://doi.org/10.5281/zenodo.3598515.
2. In this and the following figure, the font groups "other font" and "not a font" are not shown as they would mainly represent noise (images, blank pages, tables etc.).

Bibliography

Hessel, Alfred (1937): "Die Schrift der Reichskanzlei seit dem Interregnum und die Entstehung der Fraktur", in: *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen*. Philologisch-Historische Klasse. Fachgruppe 2: Nachrichten aus der Mittleren und Neueren Geschichte N. F. 2: 43-59.

Huang, Gao / Liu, Zhuang / van der Maaten, Laurens / Weinberger, Kilian Q. (2017): „Densely Connected Convolutional Networks", in: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2261-2269.

Kapr, Albert (1993): *Fraktur*. Form und Geschichte der gebrochenen Schriften. Mainz: Schmidt.

Kautzsch, Rudolf (1922): *Die Entstehung der Frakturschrift* (= Jahresbericht der Gutenberg-Gesellschaft 20, Beilage). Mainz: Gutenberg-Gesellschaft.

Kapr, Albert (1993): Fraktur. Form und Geschichte der Gebrochenen Schriften. Mainz: Hermann Schmidt.

Killius, Christina (1999): *Die Antiqua-Fraktur-Debatte um 1800 und ihre historische Herleitung* (= Mainzer Studien zur Buchwissenschaft 7). Wiesbaden: Harrassowitz.

Lühmann, Frithjof (1981): *Buchgestaltung in Deutschland 1770 - 1800*. PhD, Ludwig-Maximilians-Universität München.

Seuret, Mathias / Limbach, Saskia / Weichselbaumer, Nikolaus / Maier, Andreas / Christlein, Vincent (2019): "Dataset of Pages from Early Printed Books with Multiple Font Groups", in: *Historical Document Imaging and Processing (HIP) 2019, 5th International Workshop 21-22 September 2019, Sydney, Australia*.

Weichselbaumer, Nikolaus / Seuret, Mathias / Limbach, Saskia / Christlein, Vincent / Maier, Andreas (2019): "Automatic Font Group Recognition in Early Printed Books", in: *Digital Humanities im deutschsprachigen Raum (DHd) 2019. 6th International Conference 25-29 March 2019, Universitäten zu Mainz und Frankfurt* 84-87.

„The Vectorian" – Eine parametrisierbare Suchmaschine für intertextuelle Referenzen

Burghardt, Manuel
burghardt@informatik.uni-leipzig.de
Computational Humanities Group, Universität Leipzig

Liebl, Bernhard
Bernhard.Liebl@gmx.org
Computational Humanities Group, Universität Leipzig

Einleitung: Shakespeare, Intertextualität und computergestützte Erkennung von Zitaten

Shakespeare ist überall. Über alle zeitlichen und medialen Grenzen hinweg finden sich intertextuelle Bezüge auf die Werke von Shakespeare (vgl. Garber, 2005; Maxwell & Rumbold, 2018), der damit nicht nur der meistzitierte und meistgespielte Autor aller Zeiten, sondern auch der meistuntersuchte Autor der Welt ist (Taylor, 2016). Doch wenngleich in zahllosen Studien diverse Einzelaspekte von Shakespeares Werk aus Perspektive der Intertextualitätsforschung gründlich mittels *close reading* untersucht wurden, so gibt es bis heute keinen Überblick, kein Gesamtbild, keine systematische Karte intertextueller Shakespeare-Referenzen für größere Textkorpora. Auffällig ist zudem, dass bislang kaum Verfahren der computergestützten Erfassung intertextueller Shakespeare-Referenzen im Sinne des *distant reading* zum Einsatz kommen. Dies verwundert umso mehr, als dass sich im Bereich der Informatik und des *natural language processing* vielfältige Methoden zur Ermittlung der Ähnlichkeit zwischen Texten finden (Bär et al., 2012; Bär et al. 2015) – und nichts anderes ist Intertextualität letzten Endes. Natürlich ist hier anzumerken, dass die volle Bandbreite intertextueller Phänomene mit bloßen Mitteln der Textähnlichkeitsbestimmung nicht abgedeckt werden kann. Für unser Verständnis von Intertextualität berufen wir uns daher auf die Definition von Genette (1993) – "la présence effective d'un text dans un autre" – wobei wir unter der "effektiven Präsenz" eines Texts in einem anderen tatsächlich eine mehr oder weniger objektiv erkennbare, explizite Referenz an der Textoberfläche verstehen. Die textuelle Umschreibung einer Balkonszene mit einem Mann und einer Frau würden wir demnach nicht automatisch "Romeo and Juliet" zuordnen, was vermutlich auch nicht in allen Fällen korrekt wäre. Die folgende Variante eines bekannten Zitats aus Macbeth (Shakespeares Ursprungsvariante steht jeweils in eckigen Klammern) wäre nach unserem Verständnis hingegen objektiv aus dem Text zu erkennen und eindeutig als intertextuelle Referenz einzuordnen:

By the *stinking* [pricking] of my *nose* [thumbs], something *evil* [wicked] this way *goes* [comes]. *(Terry Pratchett: „I Shall Wear Midnight").*

Eine weitere methodische Einschränkung machen wir, indem wir Phänomene wie strukturelle Ähnlichkeit (Versmaß, Figurenkonstellation) und stilistische Ähnlichkeit[1], wie sie bspw. in der *Parodie* oder im *Pastiche* üblich sind, zunächst außer Acht lassen. In Erweiterung einer ersten Pilotstudie zur Identifizierung von Shakespearezitaten in der Fernsehserie „Dr. Who" (Burghardt et al., 2019) erproben wir in einem aktuellen Experiment das Potenzial von *word embeddings* (Mikolov et al., 2013), um so zusätzlich semantisch ähnliche oder zumindest "funktional äquivalente" (Bubenhofer, 2019) Wörter und Phrasen zu identifizieren. Durch die Auswahl unterschiedlicher *embeddings*-Modelle und weiterer, damit einhergehender Parameter (bspw. der Gewichtung anhand von Wortarten, dem Festlegen von Ähnlichkeitsschwellwerten, etc.) kann es mitunter zu sehr unterschiedlichen Ergebnissen kommen. Um hier systematisch Parameterkombinationen zu untersuchen, die möglichst optimierte Werte bzgl. *precision* und *recall* liefern, wurde im Sinne von Molnars (2019) Desiderat eines „interpretable machine learning" eine parametrisierbare Suchmaschine zur Identifizierung von Shakespeare-Referenzen als Vorstufe für einen *embeddings*-basierten Ansatz umgesetzt.

The Vectorian

Abb. 1 zeigt die Systemarchitektur der besagten Suchmaschine, die fortan als "The Vectorian"[2] bezeichnet wird. Im *Vectorian* fungieren kurze Shakespeare-Passagen (bspw. „If you prick us, do we not bleed?") als Queries; Texte, die diese Textteile (wortwörtlich oder als Variante) aufgreifen, stellen im Sinne des Information Retrieval dann die entsprechenden Ergebnisdokumente dar (für einen vergleichbaren Ansatz siehe Manjavacas et al., 2019).