

Dokument, Transkription, Forschungsdatum

Technische und kulturelle Überlegungen für interdisziplinäre Transkriptionspraxis

Baierer, Konstantin

konstantin.baierer@sbb.spk-berlin.de
Staatsbibliothek zu Berlin – Preußischer Kulturbesitz

Boenig, Matthias

boenig@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

Engl, Elisabeth

engl@hab.de
Herzog August Bibliothek Wolfenbüttel

Geestmann, Mareen

geestmann@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen

Hinrichsen, Lena

hinrichsen@hab.de
Herzog August Bibliothek Wolfenbüttel

Neudecker, Clemens

clemens.neudecker@sbb.spk-berlin.de
Staatsbibliothek zu Berlin – Preußischer Kulturbesitz

Pestov, Paul

pestov@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen

Weidling, Michelle

weidling@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen

Transkription im weitesten Sinne, d.h. die schriftliche Kodierung von Informationen, bildet die Grundlage für eine Vielzahl von Methoden in den Digital Humanities (DH). Zu transkribieren heißt, ein Quelldokument mittels Regeln und Konventionen in ein Zieldokument zu übertragen.¹ Das Dokument besteht grob betrachtet aus dem Träger, der eine Stele, Tontafel, Papyrusrolle, Papierblätter oder auch eine Datei sein kann, und dem Text.

Der Text bildet sowohl im Zusammenspiel mit dem Trägermaterial als auch in sich selbst mehrere Dimensionen aus. In den meisten Fällen soll durch die Transkription der Text erfasst werden. Die Gesamtheit der Dimensionen kann nur sehr begrenzt kodiert werden.² In Kulturerbeeinrichtungen finden sich verschiedenste Formen von Transkriptionen wie handschriftliche Kopien, Übertragungen von gesprochener Sprache in Parlamentsprotokol-

len oder auch Transkriptionen, die mit dem Ziel einer Edition erstellt werden. Darüber hinaus kann die Transkription aber auch als Eingabe für Software-Verfahren dienen. Ein Beispiel dafür wäre das Trainieren eines Neuronales Netzes (NN) für die Optical Character Recognition (OCR). Wird dieses trainierte OCR-Modell wiederum auf Dokumente angewendet, entsteht so eine automatische Transkription. Da Transkription auf verschiedenen Ebenen mit unterschiedlichen Rahmenbedingungen innerhalb der DH Anwendung findet, wollen wir uns in diesem Vortrag einigen technischen und kulturellen Aspekten dieser Praxis nähern, diese im Kontext der praktischen Transkriptionsarbeit verorten und schließlich ihren hohen Wert herausstellen.

Das Konzept der Transkription ist so alt wie die Schrift selbst. Die Kulturtechnik des Schreibens wird seit jeher gelernt durch das Abschreiben bestehender Texte (Kopie) oder das schriftliche Festhalten mündlicher Sprache (Diktat). Nur durch die kontinuierliche Transkriptionsarbeit von Kopist:innen und Übersetzer:innen im Laufe der Jahrhunderte können wir heute Texte und die ihnen zugrunde liegenden vormodernen Kulturen rezipieren. Die schriftliche Überlieferung wurde durch die Übertragungs- und Abschreibprozesse auch Veränderungen ausgesetzt, die häufig nicht dokumentiert wurden (vgl. z.B. Kammer 2017: 33–37; Schulz 2014: 290). Mit dem Aufkommen des Buchdrucks in Europa in der Renaissance kam das Setzen als neue Transkriptionsform hinzu, bei der der Aufwand für die Anfertigung von Kopien im Gegensatz zum manuellen Abschreiben mit einer steigenden Anzahl an zu produzierenden Exemplaren immer weiter sinkt. Die Folgen waren eine deutliche Ausweitung des kulturellen Austauschs, die Wiederentdeckung von bisher nur in vereinzelt Manuskripten manifesten Texten und die Ausgestaltung vieler Wissenschaftsdisziplinen wie wir sie heute kennen. Seit dem Aufkommen der Mikrochiptechnologie und des Internets befinden wir uns in der nächsten kulturtechnischen Evolution, in der IT-gestützte Verfahren auf allen Ebenen von Transkriptionsarbeit zum Einsatz kommen.

Da Informationen in IT-Systemen immer binär gespeichert werden, d.h. als Abfolge von Nullen und Einsen, stellt sich die Frage, wie und vor allem wie einheitlich Texte als Binärcode kodiert werden sollen. Bis in die 2000er Jahre hinein war ASCII, eine auf das Englische zentrierte Kodierung aus den 1960er Jahren, mit sprachspezifischen Erweiterungen (Windows-1252, KOI-8 etc.) üblich. Heute hat sich hingegen der Unicode-Standard UTF-8 weitgehend durchgesetzt, dessen Ziel es ist, für jede Letter aus jeder Schrift, historisch wie gegenwärtig, einen äquivalenten Codepoint in Unicode zu definieren. Damit ist es heute möglich, beinahe jedes Zeichen und jede Variante eindeutig zu kodieren. Es ist dementsprechend gängige Praxis, konsequent alle Texte in UTF-8 (und nicht in einer anderen, obsoleten Unicode-Codierung wie UTF-16) zu kodieren. Da Unicode ein offener und aktiv entwickelter Standard ist, werden laufend neue Zeichen darin aufgenommen. Für sehr spezielle Fälle, die (noch) nicht im Unicode-Standard vorhanden sind, gibt es Erweiterungen der Private Use Area (PUA) von Unicode, bspw. die Medieval Unicode Font Initiative (MUFI³). Die Kodierungen entsprechend MUFI sind zwar nicht Teil des offiziellen Unicode-Standards, sie sind aber dennoch gegenüber Eigenlösungen klar zu bevorzugen und werden auch in unregelmäßigen Abständen in Unicode integriert.

Neben der Schrift können in der digitalen Transkription noch weitere Bedeutung tragende (semantische) Dimensionen der Vorlage kodiert werden, bspw. Elemente des typografischen Erscheinungsbilds oder die jeweilige Hand in einem Manuskript (vgl. Neuber 2020). Der gängige Standard für die Erstellung kritischer Editionen mit reichem, semantischem Markup ist das XML-Format der Text Encoding Initiative (TEI-XML). Seit 1987 in Ent-

wicklung, ist TEI-XML ein äußerst detaillierter XML-Standard, der eine Vielzahl von Vorlagen abdeckt (Korrespondenz, Drama, Prosa, Lyrik, Handschriften uvm.). Gegenüber eigenentwickelten Kodierungen ist TEI-XML in den meisten Fällen zu bevorzugen. TEI-XML möchte die Vielzahl der textuellen Dimensionen abbilden, aus diesem Grund ist es ein sehr umfangreicher und flexibler Standard. Da die konkrete Transkription nicht alle Dimensionen erfassen soll, ist es empfehlenswert, vorhandene und in der Praxis erprobte TEI-XML-basierte Transkriptionsrichtlinien wie bspw. das DTABf (DTA-Basisformat, vgl. Haaf et al. 2014) zu nutzen. Das DTABf wird bereits seit vielen Jahren und im Austausch mit der Community gepflegt. Es bietet die Sicherheit, dass für gängige Konstrukte und textuelle Phänomene bewährte Lösungen bereits enthalten sind. Aspekte, die diese Schemas und Richtlinien nicht enthalten, können entsprechend der umfassenden TEI-Kompatibilität ergänzt werden. Dadurch kann weitgehende Interoperabilität zwischen verschiedenen Texten erreicht werden. Bei dieser Vorgehensweise sind die Transkriptionsentscheidungen bereits in bestehenden Richtlinien dokumentiert und der Transformationsaufwand der zusätzlichen textuellen Aspekte in andere Formate oder Sammlungskontexte kann minimiert werden (vgl. Fisseni et al. 2021).

Nicht immer ist das Ziel der digitalen Transkription eine kritische Edition. Moderne Verfahren der OCR und der Handwritten Text Recognition (HTR) beruhen auf Methoden des Deep Learning (DL) bzw. auf NN. Ziel dieser Verfahren ist es, die Transkription von bilddigitalisierten Werken zu automatisieren, indem der Computer lernt, Zeile für Zeile den Text im Bild zu erkennen. Diese Verfahren müssen zunächst mit manuell transkribierten Zeilenbild-Zeilentext-Paaren (Ground Truth, GT) trainiert werden, die repräsentativ für die zu erkennenden Vorlagen sind (vgl. auch Boenig et al. 2018). Während die Zielgruppe für kritische Editionen vornehmlich menschliche Rezipienten sind, werden Transkriptionen für das Training von OCR/HTR von einem Computerverfahren interpretiert und haben dadurch andere Anforderungen. Zum einen kann die semantische Auszeichnung der Transkriptionen von den derzeit implementierten Verfahren nicht verwendet werden, sie sind ausschließlich auf zeilenweisen Text fokussiert. Zum anderen ist eine durchgängig einheitliche Kodierung der GT unerlässlich, da jede Inkonsistenz das Erlernen einer inkorrekten Transkription zur Folge hat. Deshalb sind auch hier gut dokumentierte und auf den Anwendungsfall ausgerichtete Transkriptionsrichtlinien essentiell. Wir empfehlen nachdrücklich die OCR-D-Ground-Truth-Richtlinien (vgl. Boenig et al. 2019), die von OCR-D seit mehreren Jahren aktiv gepflegt werden, eine Transkription auf mehreren Ebenen der Textgenauigkeit ermöglichen und Beispiele für Sonderzeichen wie historische Ligaturen samt möglicher Normalisierungen enthalten.⁴ Diese Richtlinien nehmen Erfahrungen und Methoden auf, die in der Editionswissenschaft, in Editionsprojekten, im DFG-Projekt Deutsches Textarchiv⁵ entwickelt und gesammelt wurden⁶. Einen Schwerpunkt bildet dabei die Beschreibung und Normierung des Verhältnisses zwischen Vorlage und transkribiertem Text. Dazu wurde eine Level-Einteilung⁷ entwickelt, die den Grad der Interpretation der Übertragung in drei Levels unterteilt. Aber nicht nur zur Transkription an sich kann dieses Einteilungssystem genutzt werden, sondern auch zur Messung vorhandener Transkriptionen. So kann einfacher beurteilt werden, ob sich eine Transkription zur Verwendung als Ground Truth für das OCR-Training oder zur Evaluation eignet.

Auch in nachgelagerten Schritten der Textarbeit kommen Verfahren zum Einsatz, die als Transkriptionen angesehen werden können. Die aus Textverarbeitungsprogrammen bekannte Rechtschreibkorrektur transkribiert einen Text mit potentiellen Fehlern

zu einem korrigierten Text. Ähnlich verhält es sich mit der (semi-)automatischen Nachkorrektur von OCR mithilfe historischer Sprachmodelle. Auch die Anreicherung mit Informationen über benannte Entitäten im Text (Named Entity Recognition, NER) und deren Verknüpfung mit Normdaten (Named Entity Linking, NEL) sind Texttransformationen, deren Eingabe und Ausgabe Text mit semantischem Markup ist. Daneben gibt es eine Vielzahl von Aufgaben in der Weiterverarbeitung von Text in den diversen DH-nahen Disziplinen, die den Text anreichern, strukturieren, auswerten usw. Auch diese Aufgaben sind häufig Transkriptionen in dem Sinne, dass sich hier dieselben Fragen im Hinblick auf die Kodierung des Textes und dessen Auszeichnung mit Markup stellen. Entsprechend ist es auch hier empfehlenswert, Transkriptionsrichtlinien zu nutzen und von Anfang an explizit festzulegen, wie sich die Ausgabe eines Verfahrens im Text niederschlägt.

Schließlich sei noch auf die sich zunehmend verbreitenden Ansätze hingewiesen, die die Dichotomie von manueller und automatischer Transkription auflösen und Aspekte von beiden Vorgehensweisen kombinieren. Bspw. ist es möglich, mit einer kleinen Menge an manuell transkribierter GT zu beginnen und ein erstes, noch stark fehlerbehaftetes OCR-Modell zu trainieren. Dieses wird anschließend auf weitere Texte angewendet, deren Fehler wiederum manuell korrigiert werden. In einem iterativen Zyklus aus Training, OCR und manueller Korrektur wird das Modell schließlich so lange verbessert, bis es zufriedenstellende Ergebnisse liefert. Neben diesem Bootstrapping-Verfahren, das bspw. in OCR4all (vgl. Reul et al. 2019) Anwendung findet, gibt es auch die Möglichkeit zum Nachtrainieren. Hierbei wird ein bereits vorhandenes Modell mit zusätzlicher GT angereichert und damit bspw. für eine bestimmte Schrifttype, ein bestimmtes Werk oder einen bestimmten Zeitabschnitt angepasst. Dadurch kann mit relativ geringem Aufwand ein für die jeweilige Vorlage maßgeschneidertes OCR-Modell erstellt werden. Auch bei diesen Ansätzen ist es wichtig, nach dafür geeigneten einheitlichen Transkriptionsrichtlinien wie den OCR-D-GT-Guidelines vorzugehen, damit ein fehlerhaftes Training durch Inkonsistenzen und eine damit einhergehende schlechtere Erkennung vermieden werden.

Während die technischen Aspekte von Transkription mit fortschreitender Digitalisierung an Bedeutung gewinnen, dürfen die soziokulturellen Umstände von Transkriptionsarbeit nicht außer Acht gelassen werden. Zunächst sei darauf hingewiesen, dass Transkribieren keine neutrale Aktivität ist, sondern immer im Kontext der Transkribierenden verortet werden muss (vgl. z.B. Alpert-Abrams 2016). Bei manueller Transkription entscheidet ein Mensch, welche Phänomene – und welche nicht – in welcher Form kodiert werden sollen. Diese Selektion und Kodierungspraxis ist abhängig von der Forschungsfrage bzw. der intendierten weiteren Verarbeitung der transkribierten Daten. Während für eine stilometrische Analyse der reine Text benötigt wird, kann für andere Perspektiven auf das Werk das Layout wichtig sein. Für weitere Ansätze sind wieder andere Faktoren interessant, wie Abbildungen, handschriftlich Anmerkungen, die Materialität des Werkes betreffende Informationen wie Wasserzeichen und Papiertextur oder die Lagenformel einer Handschrift. Die Landkarte ist nicht das Gebiet – es ist unmöglich, eine Vorlage unter allen denkbaren Gesichtspunkten exakt zu beschreiben, ohne sie zu reproduzieren. Es ist aber sehr wohl möglich, diese impliziten Selektionskriterien in Form von Transkriptionsrichtlinien explizit zu machen, um potentiellen Nachnutzenden die Einschätzung der Bedeutung einer Transkription für ihre Forschungsfrage zu erleichtern.

Für die automatischen Transkriptionsverfahren gilt diese Beobachtung umso mehr, da jedes trainierte Modell nur so präzise, effektiv und umfassend sein kann wie die Trainingsdaten bzw. das

kodierte Kontextwissen, das in das Verfahren einfließt. Bei heuristischen Verfahren, in denen in Form von Software abgefasste Regeln die Verarbeitung bestimmen, sollten diese Regeln klar dokumentiert werden. Nur so kann sichergestellt werden, dass Verfahren und Daten zueinander passen. Wird bspw. ein OCR-Verfahren so trainiert, dass das "lange l" zu einem modernen "runden s" normalisiert wird, wird es niemals das "lange l" produzieren können. Für eine automatische Sprachverarbeitung könnte diese Normalisierung sogar hilfreich sein, für Forschungsfragen zu historischer Orthografie hingegen wären die Erkennungsergebnisse dann ungeeignet. Deshalb ist es auch bei automatischer Transkription wichtig, diese vorgelagerten Entscheidungen und impliziten Annahmen zu dokumentieren.

Transkription ist eine Aktivität von vielen im wissenschaftlichen Diskurs und wird daher auch von dessen Konventionen geprägt. Das oft wiederholte Mantra von "publish or perish" gilt in den DH ebenso wie in anderen Disziplinen. Mit Blick auf das Ziel, eine oder mehrere Publikationen auf Basis der Transkription zu erarbeiten, wird der eigentlichen Transkriptionsarbeit reputativ häufig kaum Bedeutung beigemessen. Insbesondere im Bereich der Informatik und der STEM-Disziplinen etabliert sich jedoch zunehmend die Ansicht, dass ein wohlkuratiertes Datenset, das gemäß der FAIR-Prinzipien⁸ in einem Forschungsdatenrepository frei nachnutzbar publiziert ist, sehr wohl eine anerkennenswerte akademische Leistung ist. Auch Drittmittelgeber wie die DFG oder die Europäische Kommission haben die Wichtigkeit von Forschungsdaten erkannt und fordern zunehmend eine Forschungsdatenstrategie als Voraussetzung für erfolgreiche Förderung. Qualitativ hochwertige Transkriptionen setzen Domänenwissen sowie größte Sorgfalt voraus und sind zeitintensiv, werden aber bislang akademisch kaum wertgeschätzt und nur selten überhaupt und wenn dann erst sehr spät veröffentlicht. Gerade im Hinblick auf die Synergienmöglichkeiten, die der rasante Fortschritt im Bereich der NN in den letzten Jahren mit sich bringt, ist es bedauerlich, wie wenig rohe Transkriptionsdaten zur Verfügung stehen, um diese Systeme zu trainieren. Ein Kulturwandel hin zu mehr Offenheit und Anerkennung für die Transkriptionsarbeit wäre somit ein doppelter Gewinn für interdisziplinäre Forschung.

Mit mehr Respekt für die Transkriptionsarbeit sollte zudem ein stärkerer, auch interdisziplinärer Austausch zur Transkriptionspraxis einhergehen, bspw. über Fragen, ob TEI-XML das beste Datenmodell für eine rohe Transkription ist oder ob nicht – wie in OCR-D (vgl. Engl et al. 2020), OCR4all, eScriptorium,⁹ Kraken¹⁰ oder Transkribus (vgl. Kahle et al. 2017) – PAGE-XML (vgl. Pletschacher und Antonacopoulos 2010) die bessere Wahl ist; welche Vor- und Nachteile alternative Kodierungsformen für verschiedene Anwendungsfälle haben können; wer in derselben Domäne bereits nach welchen Richtlinien transkribiert hat und Erfahrungen teilen kann; ob es für eine etwaige semi-automatische Erfassung mithilfe von OCR bereits passend trainierte Modelle gibt; uvm. Es ist für die DH im Speziellen, aber auch für den wissenschaftlichen Diskurs insgesamt, ein Gewinn, wenn wir eine interdisziplinäre Diskussion zur Transkriptionspraxis weiterführen. Dadurch kann es uns gelingen, reflektierter mit Ground-Truth-Daten umzugehen, bei deren Erstellung und Qualitätsbestimmung es durch fehlende Richtlinien an Transparenz und universeller Anwendbarkeit fehlt (vgl. Boenig et al. 2018).

Fußnoten

1. Vgl. Theorie der Transkription bei Sahle 2013, S. 279-283.
2. Begriff des Textes bei: Sahle 2013 und Hermes 2011.

3. <https://mufi.info>
4. <https://ocr-d.de/de/gt-guidelines/>
5. Richtlinien zur Transkription: <https://www.deutschestextarchiv.de/doku/basisformat/transkription.html>
6. Anzuführen seien auch die Dokumentationen und Richtlinien des Referenzkorpus Mittelniederdeutsch/Niederrheinisch (https://corpora.uni-hamburg.de/hzsk/de/islandora/object/file:ren-0.1_transkriptionshandbuch/datastream/PDF/transkriptionshandbuch.pdf) sowie Grundsätze für die Textbearbeitung im Fachbereich Historische Hilfswissenschaften hrsg. Archivschule Marburg Version vom 26.04.2009 (https://www.archivschule.de/uploads/Ausbildung/Grundsätze_für_die_Textbearbeitung_2009.pdf) und die Blogbeiträge zum Thema Ground Truth aus dem DFG-Projekt "Rechtsprechung im Ostseeraum" (<https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/>).
7. Übersicht der Ground-Truth-Level: <https://ocr-d.de/de/gt-guidelines/trans/trLevels.html>
8. <https://www.go-fair.org/fair-principles/>
9. <https://gitlab.com/scripta/escriptorium/>
10. <https://github.com/mittagessen/kraken>

Bibliographie

- Alpert-Abrams, Hannah** (2016): "Machine Reading the *Primeros Libros*", in: *Digital Humanities Quarterly* 10 (4) <http://www.digitalhumanities.org/dhq/vol/10/4/000268/000268.html> [letzter Zugriff 15. Juli 2021].
- Boenig, Matthias / Baierer, Konstantin / Hartmann, Volker / Federbusch, Maria / Neudecker, Clemens** (2019): "Labelling OCR Ground Truth for Usage in Repositories", in: *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage (DATECH19), Brüssel 09.05.2019* : 3-8 <https://dl.acm.org/doi/abs/10.1145/3322905.3322916> [letzter Zugriff 15. Juli 2021].
- Boenig, Matthias / Federbusch, Maria / Herrmann, Elisa / Neudecker, Clemens / Würzner, Kay-Michael** (2018): "Ground Truth: Grundwahrheit oder Ad-Hoc-Lösung? Wo stehen die Digital Humanities?", in: *DHd 2018. Kritik der digitalen Vernunft. 5. Tagung des Verbands Digital Humanities im deutschsprachigen Raum. Konferenzabstracts*. Universität zu Köln 26. Februar bis 2. März 2018: 219-223 [10.5281/zenodo.4622316](https://zenodo.org/record/4622316).
- Engl, Elisabeth / Boenig, Matthias / Baierer, Konstantin / Neudecker, Clemens / Hartmann, Volker** (2020): "Volltexte für die Frühe Neuzeit. Der Beitrag des OCR-D-Projekts zur Volltexterkennung frühneuzeitlicher Drucke", in: *Zeitschrift für Historische Forschung* 47 (2): 223-250 <https://elibrary.duncker-humboldt.com/journals/id/28/vol/47/iss/5737/art/58179/> [letzter Zugriff 15. Juli 2021].
- Fisseni, Bernhard / Sendler, Simon / Schulz, Daniela / Boenig, Matthias / Meiners, Hanna-Lena / Sikora, Uwe** (2021): "Das DTABf in der Edition – zusammenfassender Evaluationsbericht" (= CLARIAH-DE-Arbeitsberichte 1) [10.14618/ids-pub-10496](https://zenodo.org/record/10496) [im Druck].
- Haaf, Susanne / Geyken, Alexander / Wiegand, Frank** (2014): "The DTA 'base format': A TEI subset for the compilation of a large reference corpus of printed text from multiple sources", in: *Journal of the Text Encoding Initiative* 8 [10.4000/jtei.1114](https://doi.org/10.4000/jtei.1114).
- Hermes, Jürgen** (2011): *Textprozessierung - Design und Applikation*. Dissertation, Universität zu Köln <https://kups.uni-koeln.de/4561/> [letzter Zugriff 15. Juli 2021].

Kahle, Philipp / Colutto, Sebastian / Hackl, Günter / Mühlberger, Günter (2017): “Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents”, in: *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* : 19 - 24 <https://ieeexplore.ieee.org/document/8270253> [letzter Zugriff 15. Juli 2021].

Kammer, Stephan (2017): *Überlieferung: Das philologisch-antiquarische Wissen im frühen 18. Jahrhundert* (= Halle-sche Beiträge zur Europäischen Aufklärung 58). Berlin / Boston: De Gruyter 10.1515/9783110520286.

Neuber, Frederike (2020): *Das Konzept einer typografiezentrierten digitalen Edition der Werke Stefan Georges samt einem Modell zur Beschreibung von Mikrotypografie*. Dissertation, Universität zu Köln <https://kups.ub.uni-koeln.de/12202/> [letzter Zugriff 15. Juli 2021].

Pletschacher, Stefan / Antonacopoulos, Apostolos (2010): “The PAGE (Page Analysis and Ground-Truth Elements) Format Framework”, in: *Proceedings of the 20th International Conference on Pattern Recognition (ICPR2010), Istanbul, Turkey, August 23-26, 2010*. IEEE-CS Press: 257 – 260 https://www.primaresearch.org/www/assets/papers/ICPR2010_Pletschacher_PAGE.pdf [letzter Zugriff 15. Juli 2021].

Reul, Christian / Christ, Dennis / Hartelt, Alexander / Balbach, Nico / Wehner, Maximilian / Springmann, Uwe / Wick, Christoph / Grundig, Christine / Büttner, Andreas / Puppe, Frank (2019): “OCR4all — An open-source tool providing a (semi-) automatic OCR workflow for historical printings”, in: *Applied Sciences* 9 (22) 10.3390/app9224853.

Sahle, Patrick (2013): *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 3: Textbegriffe und Recodierung* (= Schriften des Instituts für Dokumentologie und Editorik 9). Dissertation, Universität zu Köln. Norderstedt: BoD <https://kups.ub.uni-koeln.de/5013/> [letzter Zugriff 15. Juli 2021].

Schulz, Matthias (2014): “Deutsch in Handschrift und gedrucktem Buch im 15. und 16. Jahrhundert”, in: Korn, Lorenz / Hoffmann, Birgitt / Stricker, Stefanie (eds.): *Aus Buchwerkstatt und Bibliothek. Manuskriptkulturen des Mittelalters in Orient und Okzident* (= Bamberger interdisziplinäre Mittelalterstudien Vorträge und Vorlesungen 3). Bamberg: University of Bamberg Press: 271-304. <https://fis.uni-bamberg.de/handle/uniba/21106> [letzter Zugriff 15. Juli 2021].